Predicting topsoil heavy metal content of parcels of land: An empirical validation of customary and constrained lognormal block kriging and conditional simulations

Predicting topsoil heavy metal content of parcels of land: An empirical validation of customary and constrained lognormal block kriging and conditional simulations[1]

Christoph Hofer[a,b], Franz Borer[c], Roland Bono[d], Achim Kayser[e], Andreas Papritz[a,*]

[a] *Institute of Terrestrial Ecosystems, ETH Zurich, Universitätstrasse 16, 8092 Zürich, Switzerland*
[b] *Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Rosenstrasse 3, 8400 Winterthur, Switzerland*
[c] *Amt für Umwelt, Werkhofstrasse 5, 4509 Solothurn, Switzerland*
[d] *Amt für Umweltschutz und Energie, Fachstelle Bodenschutz, Rheinstrasse 29, 4410 Liestal, Switzerland*
[e] *Amt für Umwelt des Kantons Thurgau, 8510 Frauenfeld, Switzerland*

Abstract

Soil contamination by heavy metals is an important problem in many countries. As a first step in mitigating the related health risks, one has to delineate zones where metal concentrations exceed tolerable levels. Predictions of metal concentrations are usually required for blocks because remediation or regulatory decisions are imposed for entire parcels. Parcel areas typically exceed the observation support, but are smaller than the survey domain. Mapping soil pollution involves therefore a local change of support. Using data from an extensive survey of heavy metals in the soils around a metal smelter, we validate in this study geostatistical block predictions with measured heavy metal concentrations that were representative for the mean metal content on 53 parcels with areas of 500–5500 m^2. Block predictions were computed by conditional simulations (CS) and several variants of lognormal universal (LUK), constrained (LCK) and covariance-matching constrained (LCMCK) block kriging from observations with quasi-point support (2–100 m^2). Lognormal block kriging predictions were either computed based on the assumption that both observations and block means are lognormally distributed or by averaging lognormal point kriging predictions. Target quantities were the block means of metal content in 0–20 cm depth and exceedance of regulatory thresholds by these means. CS gave the most precise predictions, both of block means and of threshold exceedance. However, the advantage was not pronounced: LUK, although slightly negatively biased, predicted block means nearly as well and was not much worse than LCK, LCMCK or CS when predicting threshold exceedance. LCK was partly positively biased (in particular when averaging lognormal constrained point kriging predictions) and was clearly less precise than LUK and CS when predicting block means. All four methods predicted threshold exceedance with good success as judged by the areas under Receiver Operating Characteristic curves (0.78–0.92). The good performance of LUK was rather surprising because nonlinear transforms of customary block kriging predictions are commonly

[1] NOTICE: This is the author's version of a work that was accepted for publication in Geoderma. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms, may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Geoderma, *Predicting topsoil heavy metal content of parcels of land: An empirical validation of customary and constrained lognormal block kriging and conditional simulations*, VOLUME 193–194, PAGES 200–212, 2013, DOI 10.1016/j.geoderma.2012.08.034.

known to be biased because kriging predictions are smoother than the target quantities. The relative success of LUK must be attributed to dense soil sampling around the validation parcels that dominantly lay in the severely contaminated part of the survey domain where a lot of soil samples had been taken. When sampling is dense the smoothing bias of block kriging does not matter much. In this situation, we can expect only limited gains in the precision of predictions by more sophisticated methods such as CS and LCK.

Keywords: Local change of support, Lognormal block kriging, Validation, Heavy metals, Soil contamination

1. Introduction

Soil contamination by heavy metals is a problem in many countries worldwide. Sometimes, the pollution is so severe that the health of humans and other organisms is at risk. To avert harm one has to spatially delineate the zone where the metal concentration exceeds a tolerable level and to take protective measures there. Geostatistical methods have become an important tool for mapping soil pollution at the local and regional scales (Altfelder et al., 2002; Hani and Pazira, 2011; Karanlık et al., 2011; Marchant et al., 2011; McGrath et al., 2004; Papritz et al., 2005; Paul and Cressie, 2011; Rawlins et al., 2006; Saito and Goovaerts, 2001; Sollitto et al., 2010; Xie et al., 2011). Apart from predicting the concentration per se, geostatistical methods were also used to predict whether the metal concentration exceeds regulatory thresholds that have been enacted in many countries in recent years (Chu et al., 2010; Lin et al., 2002; Marchant et al., 2011; Papritz et al., 2005).

Soil protection authorities impose protective measures usually for entire parcels of land. The metal content and threshold exceedance must therefore be predicted for small blocks of land, typically having areas of a few hundred to several thousand m^2. The support of the observations, i.e. the area over which the material of a composite soil sample is collected, is usually (much) smaller (5–100 m^2). For planning protective measures, one faces therefore a nonlinear local change of support problem (Chilès and Delfiner, 1999, pp. 435–437, Gotway and Young, 2002): Based on usually sparse quasi-point support observations, one has to predict for parcels — which are small compared to the area of the survey domain — whether their means exceed a threshold. In the sequel, we adopt the usual geostatistical terminology and use the terms block and block mean for a parcel and the mean content of the pollutant on it.

Conditional simulations (CS, Chilès and Delfiner, 1999, pp. 449–592) are usually preferred when one has to predict threshold exceedance by block means in soil pollution studies (Ersoy et al., 2008; Lin et al., 2001; Papritz et al., 2005). However, CS is highly parametric, and nonlinear predictions by CS may be badly biased if the probabilistic model is misspecified as Aldworth and Cressie (2003) pointed out. Universal block kriging (UK, Cressie, 1993b, p. 155) is simpler than CS and commonly believed to be less sensitive to misspecifications of the model. But it is well known that nonlinear functionals of UK block predictions are commonly biased because the variance of the UK predictor is smaller than the variance of the target quantity (Aldworth

*Corresponding author
Email addresses: christoph.hofer@alumni.ethz.ch (Christoph Hofer), andreas.papritz@env.ethz.ch (Andreas Papritz)

and Cressie, 2003). The constrained (CK, Cressie, 1993a) and the covariance-matching constrained kriging predictors (CMCK, Aldworth and Cressie, 2003) share with UK simplicity and robustness but are less biased than UK for nonlinear predictions and exactly unbiased for smooth nonlinear functionals of a Gaussian variable. Like UK, the CK and CMCK predictors are linear in the data, but they satisfy in addition to the usual unbiasedness constraint of UK further constraints: CMCK matches the covariance matrix of predictions and block means for a set of blocks and CK does the same but just for a single block. These additional constraints eliminate or reduce the bias of predictions of nonlinear functionals of block means.

Aldworth and Cressie (2003) and Hofer and Papritz (2010) investigated the properties of CK and CMCK in comparison to CS and UK by simulations for the situation of global (Aldworth and Cressie, 2003) and local change of support (Hofer and Papritz, 2010). In our former article we studied a scenario where a point source has polluted the soils in its neighbourhood. We considered Gaussian and positively skewed spatial processes with a nonstationary mean function and various scenarios for the auto-correlated error. For Gaussian data and blocks with locally dense sampling CS, UK, CK and CMCK performed equally well, both for predicting the block means and whether they exceed a threshold. When sampling was sparse CK and CMCK gave less precise predictions of the block means — which is expected from theory — but outperformed UK for predicting threshold exceedance, irrespective of the data distribution. CK was only outperformed by CS in the Gaussian case when threshold exceedance was predicted by the conditional quantiles. However, CS was strongly biased for the skewed data whereas CK still provided unbiased and quite precise exceedance predictions. As CMCK was not clearly better than CK, we recommended the latter method to predict block means and nonlinear transforms thereof because this method seems to offer a good compromise between robustness against model misspecification, precision of the predictions and simplicity to compute.

Unlike the validation of point predictions of soil attributes, which found some attention in the past (e.g. Bourennane et al., 2007; Brus et al., 1996; Buttafuoco et al., 2007; Laslett, 1994; Moyeed and Papritz, 2002; Papritz and Dubois, 1999; Voltz and Webster, 1990), we are not aware of any study that validates geostatistical predictions empirically for the situation of local change of support, i.e., when linear and nonlinear block predictions are computed from data observed on quasi-point support. The reason for this lack might be that soil samples with support larger than about 100 m^2 are rarely gathered in studies on the spatial variation of soil properties. Such block-support data were available from a comprehensive survey (Kayser et al., 2006; Papritz et al., 2005) of the heavy metal content of the soils in a region near Dornach, NW Switzerland, where a metal smelter had polluted the soils by copper (Cu), zinc (Zn) and cadmium (Cd). The study had been commissioned in 2003 by the owner of the factory and the soil protection authorities with the aim to delineate the zones where the trigger and clean-up thresholds of the Swiss soil protection ordinance (OIS, 1998) were exceeded and to examine whether protective measures were required in these zones.

As we believe that empirical validation studies are important to guide practitioners in their choice of an adequate spatial prediction method we compare here the performance of CS, UK, CK and CMCK empirically with data from this survey. We validate lognormal CS, UK, CK and CMCK block predictions of i) the mean topsoil metal content on 53 parcels and ii) predictions whether these means exceed regulatory thresholds. The validation is done by comparing the predictions, computed from observations with support of 2–100 m^2, with measured metal content in composite (bulked) soil samples for which the individual cores had been evenly distributed over the parcels, having areas between 400 and 5500 m^2 (mean 1300 m^2).

The sites where soil was sampled in the 2003 survey were selected based on CS predictions

of the mean topsoil Cu content of some 7400 parcels, computed with data from earlier surveys, see Papritz et al. (2005) for details. If the 95 %-quantile of the predictive distribution of the Cu block mean exceeded for a given parcel the clean-up threshold then a composite sample covering the whole unsealed part of the parcel was collected. All such parcels were sampled in this way, provided there were no measurements from earlier surveys. In the zone where the CS predictions suggested less severe contamination (95 %-quantile $<$ clean-up threshold) 'standard' soil samples were collected, obtained by bulking the cores over an area of about 100 m^2.

The contrasting support of the two types of soil samples allowed us to validate block predictions by splitting the data into two subsets: A calibration set, consisting of all samples with support ≤ 100 m^2, was used to compute the predictions, and a validation set with 53 observations relating to samples with support > 500 m^2 served us to compare the precision of the CS, UK, CK and CMCK predictions. We report here the results of this comparison.

The remainder of the article is structured as follows: Section 2 summarises the essential theory about conditional simulations, lognormal universal and constrained block kriging. In the following section, we describe the study site (3.1), the available data (3.2), the model fitted to the calibration data (3.3), the target quantities (3.4) and the criteria used to validate the predictions (3.5). A subsection on computations (3.6) completes section 3. Sections 4 and 5 present and discuss the results and the article concludes with some final remarks in section 6.

2. Conditional simulations, lognormal universal, constrained and covariance-matching constrained block kriging

Let $\mathbf{Z} = (Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n))'$ denote the vector of random variables that model the observations $z(\mathbf{s}_i), i = 1, 2, \ldots, n$ ($'$ denotes transpose). We assume a lognormal distribution for $Z(\mathbf{s})$:

$$Z(\mathbf{s}) = \exp\big(Y(\mathbf{s})\big) = \exp\big(S(\mathbf{s}) + \varepsilon(\mathbf{s})\big) = \exp\big(\mathbf{x}(\mathbf{s})'\beta + \delta(\mathbf{s})\big) \cdot \exp\big(\varepsilon(\mathbf{s})\big), \tag{1}$$

where $S(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\beta + \delta(\mathbf{s})$ is a Gaussian variable with expectation $\mathrm{E}[S(\mathbf{s})] = \mathbf{x}(\mathbf{s})'\beta$; $\mathbf{x}(\mathbf{s})$ and β are p-vectors with the covariates for location $\mathbf{s}$ and the regression coefficients; $\delta(\mathbf{s})$ is a zero mean weakly stationary Gaussian variable with isotropic covariance function $\mathrm{Cov}[\delta(\mathbf{s}), \delta(\mathbf{s}+\mathbf{h})] = \mathrm{Cov}[S(\mathbf{s}), S(\mathbf{s}+\mathbf{h})] = C(h)$, $h = ||\mathbf{h}||$; and $\varepsilon(\mathbf{s})$ is a Gaussian zero mean white noise variable with variance σ^2_ε that models independent measurement errors. Notice that the errors are multiplicative on the original scale of the measurements.

Our task is to predict the block mean of the measurement error-free variable $W(\mathbf{s}) = \exp\big(S(\mathbf{s})\big)$

$$W(B) = \frac{1}{|B|} \int_B \exp\big(S(\mathbf{s})\big)\, \mathrm{d}\mathbf{s}, \tag{2}$$

and to predict whether $W(B)$ exceeds some threshold T ($|B|$ denotes the area of the block). A standard approach is to use conditional simulations to predict $W(B)$: A conditional realisation of $W(B)$ given $\mathbf{Z}$, say $W_\omega(B)|\mathbf{Z}$, is obtained from

$$W_\omega(B)|\mathbf{Z} = \frac{1}{|B|} \int_B \exp\big(S_\omega(\mathbf{s})|\mathbf{Y}\big)\, \mathrm{d}\mathbf{s},$$

where $S_\omega(\mathbf{s})|\mathbf{Y}$ is a realisation of the Gaussian random process $\{S(\mathbf{s})\}$ over the domain of interest,

conditioned to $\mathbf{Y} = \log\big(\mathbf{Z}\big)$. In practice, we approximate the above integral by the sum

$$W_\omega(B)|\mathbf{Z} \approx \frac{1}{n_B} \sum_{\mathbf{s}_i \in B} \exp\big(S_\omega(\mathbf{s}_i)|\mathbf{Y}\big), \tag{3}$$

where n_B is the number of points — usually arranged on a grid — that fall into the block. Such conditional Gaussian realisations can be efficiently simulated by the 'conditioning-by-kriging' method (Chilès and Delfiner, 1999, pp. 466). The best mean square predictor of $W(B)$ is the conditional expectation $\mathrm{E}[W(B)|\mathbf{Z}]$. We can approximate this quantity numerically by simulating a large number M of realisations of $W_\omega(B)|\mathbf{Z}$ and by computing

$$\widehat{W}_{\mathrm{CS}}(B) = \frac{1}{M} \sum_{\omega=1}^{M} W_\omega(B)|\mathbf{Z}. \tag{4}$$

The lognormal block kriging predictor computed under the so-called assumption of permanence of lognormality is another predictor of (2) that was used a lot in the past (cf. Cressie, 2006, for references). This predictor is obtained by predicting the block mean, $S(B) = 1/|B| \int_B S(\mathbf{s})\,\mathrm{d}\mathbf{s}$, of the log-transformed, error-free variable $S(\mathbf{s})$ linearly by block kriging and by transforming the prediction back to the original scale of the measurements with a bias adjustment that is computed under the assumption that both $\exp(S(\mathbf{s}))$ and $\exp(S(B))$ are lognormally distributed, which strictly cannot hold.

Before we consider this adjustment, we review different ways to predict $S(B)$: Universal kriging provides the best linear unbiased plug-in predictor

$$\widehat{S}_{\mathrm{UK}}(B) = \mathbf{x}(B)'\widehat{\beta}_{\mathrm{GLS}} + \mathbf{c}(\mathbf{s}_{1\ldots n}, B)'\,\Sigma^{-1}\,(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{GLS}}), \tag{5}$$

where $\mathbf{x}(B)$ is the p-vector with the covariates for the target block B; $\widehat{\beta}_{\mathrm{GLS}}$ is the p-vector with the generalised least squares estimate of β; $\mathbf{c}(\mathbf{s}_{1\ldots n}, B)$ is an n-vector with the covariances between $\mathbf{Y}$ and $S(B)$; $\Sigma = \mathrm{Cov}[\mathbf{Y}, \mathbf{Y}']$ is the $n \times n$ covariance matrix of $\mathbf{Y}$; and $\mathbf{X}$ is the $n \times p$ design matrix of the data. The mean squared prediction error (MSPE) of UK is given by

$$\begin{aligned} \mathrm{MSPE}[\widehat{S}_{\mathrm{UK}}(B)] &= \mathrm{Var}[S(B) - \widehat{S}_{\mathrm{UK}}(B)] \\ &= \mathrm{Var}[S(B)] - \mathbf{c}(\mathbf{s}_{1\ldots n}, B)'\,\Sigma^{-1}\,\mathbf{c}(\mathbf{s}_{1\ldots n}, B) + \\ &\quad \big(\mathbf{x}(B) - \mathbf{X}'\Sigma^{-1}\mathbf{c}(\mathbf{s}_{1\ldots n}, B)\big)'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\big(\mathbf{x}(B) - \mathbf{X}'\Sigma^{-1}\mathbf{c}(\mathbf{s}_{1\ldots n}, B)\big). \end{aligned} \tag{6}$$

It is well-known that for any nonlinear function $g(\cdot)$, $g\big(\widehat{S}_{\mathrm{UK}}(B)\big)$ is a biased predictor of $g\big(S(B)\big)$ as $\widehat{S}_{\mathrm{UK}}(B)$ underestimates the true variation of $S(B)$. Cressie (1993a) and Aldworth and Cressie (2003) therefore proposed linear predictors that are less biased in these circumstances because they match either for a set of blocks (CMCK) or only a single block (CK) the (co-)variances of predictions and block means. The CMCK predictor, $\widehat{\mathbf{S}}_{\mathrm{CMCK}}$, of Aldworth and Cressie of a set of m block means, $\mathbf{S} = (S(B_1), \ldots, S(B_m))'$, is given by

$$\widehat{\mathbf{S}}_{\mathrm{CMCK}} = \mathbf{X}_B\widehat{\beta}_{\mathrm{GLS}} + \mathbf{K}'\mathbf{C}'\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{GLS}}), \tag{7}$$

where $\mathbf{K} = \mathbf{Q}_1^{-1}\mathbf{P}_1$ is an $m \times m$ matrix; $\mathbf{Q}_1$ and $\mathbf{P}_1$ are symmetric $m \times m$ matrices given by

$$\mathbf{Q}_1\mathbf{Q}_1 = \mathbf{Q} = \mathbf{C}'(\Sigma^{-1} - \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1})\mathbf{C}, \tag{8}$$

$$\mathbf{P}_1\mathbf{P}_1 = \mathbf{P} = \Sigma_B - \mathbf{X}_B(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}_B', \tag{9}$$

where $\mathbf{C} = (\mathbf{c}(\mathbf{s}_{1\ldots n}, B_1), \ldots, \mathbf{c}(\mathbf{s}_{1\ldots n}, B_m))$ is an $n \times m$ matrix with covariances between the observations and the block means; $\Sigma_B = \mathrm{Cov}[\mathbf{S}, \mathbf{S}']$ is the $m \times x$ covariance matrix of $\mathbf{S}$ and $\mathbf{X}_B = (\mathbf{x}(B_1), \ldots, \mathbf{x}(B_m))'$ is the $m \times p$ design matrix of the m blocks. The matrices $\mathbf{Q}_1$ and $\mathbf{P}_1$ exist and are positive definite if $\mathbf{Q}$ and $\mathbf{P}$ are themselves positive definite. In practice, the CMCK predictor thus exists if $\mathbf{Q}$ and $\mathbf{P}$ are positive definite. Unlike $\mathbf{Q}$, which is always nonnegative definite, the matrix $\mathbf{P}$ may become negative definite (see Aldworth and Cressie, 2003, p. 15, and Hofer and Papritz, 2010, p. 635 for details).

For $m = 1$ the CMCK predictor simplifies to the CK predictor, first proposed by Cressie (1993a)

$$\widehat{S}_{\mathrm{CK}}(B) = \mathbf{x}(B)'\widehat{\beta}_{\mathrm{GLS}} + K\,\mathbf{c}(\mathbf{s}_{1\ldots n}, B)'\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathrm{GLS}}), \tag{10}$$

where the scalar K is given by

$$K = \left(\frac{P}{Q}\right)^{1/2} = \left(\frac{\mathrm{Var}[S(B)] - \mathbf{x}(B)'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{x}(B)}{\mathbf{c}(\mathbf{s}_{1\ldots n}, B)'(\Sigma^{-1} - \Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1})\,\mathbf{c}(\mathbf{s}_{1\ldots n}, B)}\right)^{1/2}. \tag{11}$$

The CK predictor exists if the numerator and denominator of (11) are positive. Whereas $Q \geq 0$, P may become negative, which is more likely to happen if $|B|$ is large ($\mathrm{Var}[S(B)]$ small) or if the trend is extrapolated ($\mathbf{x}(B)'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{x}(B) = \mathrm{Var}[\mathbf{x}(B)'\widehat{\beta}_{\mathrm{GLS}}]$ large). Hofer and Papritz (2010, pp. 643–645) discuss in more detail what controls P and Q.

The constraint that the (co-)variances of $\widehat{\mathbf{S}}$ and $\mathbf{S}$ must match, results in larger mean squared errors of the CK and CMCK compared to the UK predictor

$$\mathrm{MSPE}[\widehat{\mathbf{S}}_{\mathrm{CMCK}}] = \mathrm{MSPE}[\widehat{\mathbf{S}}_{\mathrm{UK}}] + (\mathbf{P}_1 - \mathbf{Q}_1)(\mathbf{P}_1 - \mathbf{Q}_1), \tag{12}$$

where

$$\begin{aligned} \mathrm{MSPE}[\widehat{\mathbf{S}}_{\mathrm{UK}}] &= \mathrm{Cov}[(\mathbf{S} - \widehat{\mathbf{S}}_{\mathrm{UK}}), (\mathbf{S} - \widehat{\mathbf{S}}_{\mathrm{UK}})'] \\ &= \Sigma_B - \mathbf{C}'\Sigma^{-1}\mathbf{C} + (\mathbf{X}_B' - \mathbf{X}'\Sigma^{-1}\mathbf{C})'(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}_B' - \mathbf{X}'\Sigma^{-1}\mathbf{C}), \end{aligned} \tag{13}$$

is the covariance matrix of the UK prediction errors. Similarly, the mean squared error of CK is given by

$$\mathrm{MSPE}[\widehat{S}_{\mathrm{CK}}(B)] = \mathrm{MSPE}[\widehat{S}_{\mathrm{UK}}(B)] + (\sqrt{P} - \sqrt{Q})^2. \tag{14}$$

To transform the predictions $\widehat{S}_{\ldots}(B)$ back to the original scale under the assumption of permanence of lognormality, one uses the approximately unbiased back-transformation (Cressie, 2006)

$$\widehat{W}_{\mathrm{L}\ldots}(B) = \exp\left(\widehat{S}_{\ldots}(B) + 1/2\left\{\mathrm{Var}[\mathrm{S}(\mathbf{s})] + \beta'\mathbf{M}(B)\,\beta - \mathrm{Var}[\widehat{S}_{\ldots}(B)]\right\}\right), \tag{15}$$

where $_{...}$ stands for UK, CK or CMCK and

$$\mathbf{M}(B) = \frac{1}{|B|}\int_B \big(\mathbf{x}(\mathbf{s}) - \mathbf{x}(B)\big)\big(\mathbf{x}(\mathbf{s}) - \mathbf{x}(B)\big)' \,\mathrm{d}\mathbf{s}$$

is the (spatial) covariance matrix of the covariates for a point $\mathbf{s}$ uniformly distributed in B. To compute (15) we thus needed $\mathrm{Var}[\widehat{S}_{\mathrm{CMCK}}(B)]$, $\mathrm{Var}[\widehat{S}_{\mathrm{CK}}(B)]$ and $\mathrm{Var}[\widehat{S}_{\mathrm{UK}}(B)]$, which are given by

$$\mathrm{Var}[\widehat{S}_{\mathrm{CKCK}}(B)] = \mathrm{Var}[\widehat{S}_{\mathrm{CK}}(B)] \quad = \quad \mathrm{Var}[S(B)] = \frac{1}{|B|^2}\int_B\int_B C(||\mathbf{s}-\mathbf{t}||)\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{t} \tag{16}$$

$$\mathrm{Var}[\widehat{S}_{\mathrm{UK}}(B)] \quad = \quad \mathrm{Var}[S(B)] - \mathrm{MSPE}[\widehat{S}_{\mathrm{UK}}(B)] + 2\,\Psi(B)'\mathbf{x}(B), \tag{17}$$

where

$$\Psi(B) = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\big(\mathbf{x}(B) - \mathbf{X}'\Sigma^{-1}\mathbf{c}(\mathbf{s}_{1\ldots,n}, B)\big)$$

is the vector with the p Lagrange multipliers of the UK equations.

Cressie (2006) showed by simulation that (15) may be biased and less efficient than the predictor

$$\widetilde{W}_{\mathrm{L}...}(B) = \frac{1}{|B|}\int_B \widehat{W}_{\mathrm{L}...}(\mathbf{s})\,\mathrm{d}\mathbf{s},$$

which we approximate again by a sum

$$\widetilde{W}_{\mathrm{L}...}(B) \approx \frac{1}{n_B}\sum_{\mathbf{s}_i \in B} \widehat{W}_{\mathrm{L}...}(\mathbf{s}_i), \tag{18}$$

where

$$\widehat{W}_{\mathrm{L}...}(\mathbf{s}_i) = \exp\big(\widehat{S}_{...}(\mathbf{s}_i) + 1/2\{\mathrm{Var}[S(\mathbf{s}_i)] - \mathrm{Var}[\widehat{S}_{...}(\mathbf{s}_i)]\}\big) \tag{19}$$

are unbiasedly back-transformed lognormal point kriging predictions. Cressie's (2006) simulation results suggest that for lognormal ordinary kriging at least, the bias of (15) is negligible and the loss of efficiency of (15) relative to (18) remains small, provided $\mathrm{Var}[S(\mathbf{s})]$ is not much larger than 0.1 and the dimension of the block is several times smaller than the effective range of the variogram. As both conditions were satisfied in our study, we assumed permanence of lognormality and used (15) with UK and CK predictions of $S(B)$ to predict $W(B)$. However, based upon a reviewer's suggestion, we also used (18) along with UK, CK and CMCK point predictions of $S(\mathbf{s})$. Paul and Cressie (2011) used this predictor with UK as well to predict the mean americium content of remediable blocks on a nuclear weapons test site in the United States.

3. Material and methods

3.1. Study site

The study area is situated at the northern edge of the Swiss Jura mountains, 10 km south of the city of Basel (7° 37' E, 47° 29' N, 290–700 m above sea level). The metal smelter, which has been producing non-ferrous metal alloys and semi-finished products since 1895, polluted the soils by dust emissions until the 1980s in four villages (Fig. 1): Dornach, situated on a west facing slope 600 m east of the smelter, was most strongly affected; Arlesheim (on the same slope further to the north outside of the map section), Aesch and Reinach, located on the alluvial plane in the SW and NW, less so. The premises of the smelter border residential areas in the sector

NNE over E to S and industrial premises in the sector from S to NW (Fig. 1). A major part of the affected area is nowadays built-up. The built-up areas expanded in Dornach and Arlesheim mainly before and in Reinach and Aesch after 1960. Land is left for agriculture on the western plain and on the hills in the (south)east, but most hills carry forests.

Jurassic limestones and marls (Bajocian, Bathonian, Oxfordian) form the bedrock on the hills in the (south)east (Bitterli-Brunner et al., 1984). These sediments are partly naturally rich in Cd. In the central and western parts of the area Tertiary (Rupelian, Chattian) or Quaternary sediments (loess, gravel) prevail. Except for the loess areas, the soils are rich in carbonate, have a high pH and are mostly sandy to silty loams. Rendzic Leptosols and calcaric Cambisols originally predominate, but at many sites the layering is disturbed by building, and artefacts (spall of bricks, etc.) are often found. Because of the high pH the soils strongly adsorb the metals, and the fraction that is readily available to plants is small (Schulin and Gupta, 2002).

3.2. Data

The data that we analysed in our study consisted of topsoil (0–20 cm) measurements of Cd, Cu and Zn at 707 (Zn: 685) sites, spread over an area of about 12 km^2 around the metal smelter. The measurements refer to composite soil samples for which several (usually 20–35) soil cores were collected over support areas of varying sizes (2–5500 m^2). Estimates of the support of the composite samples were available with some exceptions.

We split the data set into a calibration and validation set based on the support of the measurements: Measurements with support $\leq$ 100 m^2 were assigned to the calibration set (Cd and Cu: 578; Zn: 556 observations) and measurements with support $>$ 500 m^2, in total 53, formed the validation set. Measurements with support between 100 and 500 m^2 and 76 measurements without information on support were excluded from the analysis. When collecting the soil samples, a support $>$ 100 m^2 was chosen only if severe pollution was anticipated for the respective parcel (cf. introduction). Hence, our procedure to split the data into calibration and validation sets was not random because strongly polluted parcels had a bigger chance to be assigned to the validation set than parcels with minor soil pollution.

Figure 1 shows about 85 % of the calibration sites (marked by $+$). The remaining calibration sites lay outside of the displayed region, mainly to the north and east of it. The histograms of the calibration data (Fig. 2) reveal that the frequency distributions of the measurements were positively skewed. The 53 parcels to which the validation data refer are displayed in Figure 1 by red polygons. They were the targets for which predictions were computed from the calibration data. Note that the area of the validation parcels was in general larger than the recorded support of the measurements because on most parcels a part of the ground had been built-up or had been sealed in some other way.

In addition to the metal measurements, geo-referenced areal views of the study area as well as a digital cadastre, digitised geological maps and maps of changes in the land use between 1877 and 2000 were available for the geostatistical analyses.

3.3. Spatial modelling

3.3.1. Large-scale spatial trend

To account for the large-scale spatial trend, we modelled the log-transformed Cd, Cu and Zn measurements by linear regressions. An exploratory analysis of the data (not shown) revealed that the soil concentration of all three metals decreased with increasing distance, $d(\mathbf{s})$, between

a site $\mathbf{s}$ and the smelter. Besides the distance, the orientation of a site relative to the smelter mattered. We used eight 'hat' functions (B-splines of order 2, cf. Schuhmaker, 1981, sec. 4.3)

$$f_k(\mathbf{s}) = \max\left(0, 1 - \left|\frac{\theta(\mathbf{s})}{\pi/4} - k\right|\right), k = 0, 1, \ldots, 7$$

and their products with $\log(d(\mathbf{s}))$ to model a continuous change of the concentration with $d(\mathbf{s})$ and the azimuth, $\theta(\mathbf{s})$, of a site (in radian from north) as seen from the smelter. As $\sum_{k=0}^{7} f_k(\mathbf{s}) = 1$ we did not need to include an intercept, nor did we need $\log(d(\mathbf{s}))$ as a covariate on its own in the regression models.

We used in addition the categorial covariate (factor) 'Geology' since the Bajocian and Oxfordian rocks of the Swiss Jura often have elevated Cd content (Rambeau et al., 2002), which may lead to a natural enrichment of Cd in the soils by weathering. We simplified the classification of the sites with respect to their bedrock and distinguished only four groups: 'Lower/Middle Oxfordian', 'Bajocian/Upper Oxfordian', 'Lower Rupelian' and 'other bedrock' (Bathonian, Chattian, Quaternary). The former three groups have parent material rich in Cd. Zones where they outcrop are hatched with coloured lines in Figure 1. We further considered land use and its change since 1877 as a factor in the regression models. This information was also simplified: We distinguished between agricultural land/forests and areas built-up before or after 1960. The exploratory analysis further revealed that the Cd measurements of one survey (Grünenfelder and Schmidli, 1998) were systematically too large. To account for this analytical bias we included a binary variable in the regression model for $\log(\mathrm{Cd})$, which was equal to one for measurements of that survey and equal to zero otherwise.

The regression models were fitted robustly by an MM-estimator (Maronna et al., 2006, sec. 5.5). We fitted the 'full' model that contained all the covariates to the log-transformed metal data of the calibration set and simplified the models manually based on partial residual plots (Faraway, 2004, sec. 4.3). To assess the goodness of the fit of the models we used customary residual diagnostic plots (Tukey-Anscombe plot, normal Q-Q-plots, scatterplot of residuals vs. leverage, etc.). $\log(Cd(\mathbf{s}))$ was eventually modelled with the covariates $f_k(\mathbf{s})$, $f_k(\mathbf{s}) \cdot \log(d(s))$ and the factors for the simplified geology and land use as well as the indicator variable for the biased measurements. The regression models for $\log(\mathrm{Cu}(\mathbf{s}))$ and $\log(\mathrm{Zn}(\mathbf{s}))$ included the same covariates except the factor for geology and the indicator for the biased Cd measurements. Table 1 lists the GLS estimates of the regression coefficients for the three heavy metals.

Two remarks might be helpful: First, the coefficients for $f_k(\mathbf{s})$ and $f_k(\mathbf{s})\log(d(s))$ correspond to intercept and slope of a regression of $\log(\mathrm{conc}(\mathbf{s}))$ on $\log(d(\mathbf{s}))$ in the kth direction. Second, we used the so-called treatment contrasts for the factors. The coefficients listed for the various levels of the factors are therefore differences between the intercept of the respective level and the intercept of the reference level 'other bedrock' (geology) and 'agriculture/forest' (land use), respectively.

3.3.2. Variogram estimation

The sample variograms of the residuals of the robustly fitted regression models were estimated by Dowd's MAD estimator (Dowd, 1984). Then we fitted stable variogram models with a nugget component

$$\gamma(h) = \sigma_0^2\big(1 - I(h = 0)\big) + \sigma_1^2\big(1 - \exp(-(\frac{h}{\alpha})^{\kappa})\big)$$

to each sample variogram by weighted least squares (Cressie, 1993b, sec. 2.6.2). In the above equation, σ_0^2 is the nugget; $I(x)$ is an indicator function with $I(x) = 1$ if x is true and $I(x) = 0$ otherwise; σ_1^2 is the partial sill; α is the range parameter and κ is the roughness parameter. We did not try to take the non-constant support (2–100 m^2) of the calibration data into account when estimating the variograms.

The fitted variogram parameters are listed in Table 2, and the sample variograms are plotted in Figure 3, along with the fitted models. Note that we assumed $\sigma_\varepsilon^2 = 0$ for the variance of the measurement errors. Duplicated chemical analyses of a subset of the soil samples showed that $\sigma_\varepsilon^2 \approx 0.0005 - 0.002$ for the log-transformed measurements, which was several times smaller than the semivariance for the characteristic length of the measurement support (≈ 10 m). The nugget reflects therefore mostly micro-scale variation so that the assumption $W(B) \approx Z(B)$ that we used in the validation is justified.

3.4. Target quantities

We predicted for each parcel of the validation set the same quantities as in our previous simulation study (Hofer and Papritz, 2010):

1. The block mean, $W(B_i)$.
2. The binary indicator, $I(W(B_i) > T)$, that indicates if the block mean $W(B_i)$ exceeds a given threshold T.

Note that B_i refers here to the whole parcel as defined in the cadastre. We mentioned above that the recorded support of the measurements was for most parcels smaller than $|B_i|$. However, as we only had an estimate of the size of the sampled area and did not know its exact geometry, we could but predict the mean over the entire parcel.

3.5. Validation criteria

3.5.1. Block means

To validate the precision of the predictions of the block means we calculated the empirical bias (eBIAS) and the empirical mean squared prediction error (eMSPE)

$$\text{eBIAS}_k = \frac{1}{53}\sum_{i=1}^{53}(\widehat{W}_k(B_i) - W(B_i)), \tag{20}$$

$$\text{eMSPE}_k = \frac{1}{53}\sum_{i=1}^{53}(\widehat{W}_k(B_i) - W(B_i))^2, \tag{21}$$

where $\widehat{W}_k(B_i)$ is the prediction of the block mean by method k and $W(B_i)$ is the heavy metal concentration measured for block B_i (note that we assumed $\varepsilon(\mathbf{s}) = 0$).

3.5.2. Threshold exceedance

To validate predictions of threshold exceedance we adopted the approach of Hofer and Papritz (2010, pp. 640–642) and assessed the quality of the predictions for multiple thresholds. In more detail, we used for each heavy metal the ordered validation data as thresholds, say $T_{[l]}$, $l =$

$1,\ldots,53$, and computed for method k the following quantities

$$\begin{aligned}
\mathrm{TN}_{kl} &= \sum_{i=1}^{53} I(\widehat{W}_k(B_i) \le T_{[l]}) \cdot I(W(B_i)) \le T_{[l]}),\\
\mathrm{FN}_{kl} &= \sum_{i=1}^{53} I(\widehat{W}_k(B_i) \le T_{[l]}) \cdot I(W(B_i)) > T_{[l]}),\\
\mathrm{FP}_{kl} &= \sum_{i=1}^{53} I(\widehat{W}_k(B_i) > T_{[l]}) \cdot I(W(B_i)) \le T_{[l]}),\\
\mathrm{TP}_{kl} &= \sum_{i=1}^{53} I(\widehat{W}_k(B_i) > T_{[l]}) \cdot I(W(B_i)) > T_{[l]}).
\end{aligned}$$

From the counts $\mathrm{TN}_{kl},\ldots,\mathrm{TP}_{kl}$ we calculated then for each threshold and each method the Peirce skill score (PSS, Peirce, 1884). PSS characterises the conditional bias and equals the difference between the hit, H_{kl}, and the false alarm rate, F_{kl}

$$\mathrm{PSS}_{kl} = \mathrm{H}_{kl} - \mathrm{F}_{kl}, \tag{22}$$

where

$$\mathrm{H}_{kl} = \frac{\mathrm{TP}_{kl}}{\mathrm{FN}_{kl} + \mathrm{TP}_{kl}}, \tag{23}$$

and

$$\mathrm{F}_{kl} = \frac{\mathrm{FP}_{kl}}{\mathrm{TN}_{kl} + \mathrm{FP}_{kl}}. \tag{24}$$

PSS ranges between -1 (perfect misclassification: $\mathrm{TN}_{kl} = \mathrm{TP}_{kl} = 0$) and 1 (perfect classification: $\mathrm{FN}_{kl} = \mathrm{FP}_{kl} = 0$). Tossing a fair coin results in $\mathrm{PSS} = 0$.

To see whether and how strongly the kth method systematically over- or underestimated the exceedance frequency we used in addition the bias score (BS, Wilks, 2011, p. 310)

$$\mathrm{BS}_{kl} = \frac{\mathrm{FP}_{kl} + \mathrm{TP}_{kl}}{\mathrm{FN}_{kl} + \mathrm{TP}_{kl}}. \tag{25}$$

$\mathrm{BS}_{kl} = 1$ indicates that method k predicts on average the correct number of threshold exceedances, whereas $\mathrm{BS}_{kl} > 1$ ($\mathrm{BS}_{kl} < 1$) signals systematic overestimation (underestimation) of exceedance. More information on PSS and BS may be found in Hofer and Papritz (2010) and in the references cited therein.

3.6. Computations

We used the software R (R Development Core Team, 2012) for all the computations. Robust regressions were computed by the function lmrob of the R package robustbase (Rousseeuw et al., 2011), and the variogram models were evaluated by the respective function of the R package RandomFields (Schlather, 2001, 2011).

Block kriging predictions (UK, CK) of the log-transformed metal content were computed by the preCKrige and CKrige functions of the R package constrainedKriging (Hofer and Papritz, 2011) and were transformed back to the original scale under the assumption of permanence of lognormality by equations (15), (16) and (17). We denote the respective predictions in the sequel

by LUKp and LCKp. In addition, we computed point UK, CK and CMCK predictions of $S(\mathbf{s})$ on a 10×10 m^2 grid. One grid point thus represented an area of 100 m^2, equal to the support of the vast majority of calibration measurements. To constrain the CMCK predictions, we used 3×3 sets of points centred on the target point in the middle. The UK, CK and CMCK point predictions were transformed back by (19), and we computed the means for all sets of points falling into one of the 53 validation parcels. These predictions are denoted by LUK, LCK and LCMCK respectively. Note that we did not use local search windows in kriging, but computed the predictions always from all the observations in the calibration set.

Gaussian conditional simulations were generated by the kriging method (Chilès and Delfiner, 1999, pp. 465–472). To this end, unconditional Gaussian realisations of $\{\delta(\mathbf{s})\}$ were simulated at the 611×611 nodes of a 10×10 m^2 grid by the circulant embedding algorithm of Chan and Wood (1997), implemented in the function GaussRF of the R package RandomFields. For each heavy metal, we simulated 2000 realisations of conditional errors and added the GLS trend surface to obtain 2000 conditional realisations of the log-transformed metal content at all the grid nodes. The simulated values were then transformed back to the original scale, and 2000 conditional realisations of the block means were obtained by arithmetically averaging the grid values within the 53 validation parcels. The block means were then predicted by the means of the conditional realisations (conditional means) and exceedance of threshold $T_{[l]}$ by the $\frac{l}{53}$-quantile of the 2000 conditionally simulated block means. This choice should maximise PSS (Hofer and Papritz, 2010, p. 649).

Apart from the conditional mean (cf. equation 4) and the various kriging predictions $\widehat{W}_{\mathrm{LUKp}}(B)$, $\widehat{W}_{\mathrm{LCKp}}(B)$, $\widehat{W}_{\mathrm{LUK}}(B)$, $\widehat{W}_{\mathrm{LCK}}(B)$, and $\widehat{W}_{\mathrm{LCMCK}}(B)$ we used as a last predictor the bias-corrected, back-transformed GLS trend surface predictor

$$\widehat{W}_{\mathrm{GLS}}(B) = \exp\left(\mathbf{x}(B)'\widehat{\beta}_{\mathrm{GLS}} + 1/2\left\{\mathrm{Var}[\mathrm{S}(\mathbf{s})] + \widehat{\beta}'_{\mathrm{GLS}}\,\mathbf{M}(B)\,\widehat{\beta}_{\mathrm{GLS}} - \mathrm{Var}[\mathbf{x}(B)'\widehat{\beta}_{\mathrm{GLS}}]\right\}\right) \quad (26)$$

to predict $W(B)$.

4. Results

The LCKp prediction of the Cd content of one parcel west of the smelter (marked by $*$ in Fig. 1) did not exist because P was negative. We excluded the respective datum therefore from the Cd validation set.

4.1. Validating predictions of parcel means

The GLS, CS and LCKp predictions are plotted in Figure 4 against the measured contents of the three metals. The line segments in the plots represent differences in predictions by the respective method and LUKp, which was used as baseline. Figures 5 and 6 report empirical biases (eBIAS) of predictions and mean squared prediction errors (eMSPE) for all the methods.

The GLS trend surface predictions were conditionally biased: Small contents were over- and large content underestimated (Fig. 4A–C). The other methods showed no or lesser conditional biases (e.g. CS and LCKp in Fig. 4D–I). The CS, LUKp and LUK predictions (latter not shown in Fig. 4) hardly differed from each other. Compared to LUKp, the LCKp, LCK and LCMCK predictions were 'amplified' in the sense that small predictions were smaller and large predictions larger than the respective LUKp predictions ($K > 1$, cf. equation 10). On the log-scale

the discrepancies between LUKp and the constrained predictions were approximately symmetric (Fig. 4J–L), but they became asymmetric by the back-transformation (Fig. 4G–I for LCKp).

For Cd the largest marginal biases were observed for LUK and LCK, and LCMCK was best, closely followed by CS (Fig. 5). For Cd, the contributions of eBIAS to eMSPE remained small ($<$ 2 % for LUK and LCK, Fig. 6). This was different for Cu and Zn where large eBIASs contributed noticeably to eMSPEs for GLS, LCK and LCMCK. For all the metals, CS had consistently the smallest eMSPE (Fig. 6), and LCK performed consistently worst (eMSPEs about twice as large as those of other methods). Strong amplification of the largest LCK predictions relative to LUK was the reason for the poor precision. The stronger amplification of large LCK compared to large LCKp predictions was caused by larger values of $\sqrt{P}$ for point than for block predictions on the log-scale ($\mathrm{Var}[S(\mathbf{s})] \gg \mathrm{Var}[S(B)]$, cf. equation 11). In general, the eMSPEs of CS, LUKp and LUK did not differ much, and for Cd those of LCKp and LCMCK were not much worse, unlike Cu and Zn where all constrained predictors were clearly less precise than conditional simulations or either type of lognormal universal block kriging.

4.2. Validating predictions of threshold exceedance

The Peirce skill scores (PSS) of CS, LUKp, LCKp and LCMCK are plotted in Figures 7A–C for the three heavy metals against the 53 (Cd: 52) thresholds. The PSS statistic, being the difference between the hit and false alarm rate, characterises the conditional bias of binary ('yes'/'no') predictions. The larger PSS, the better the classification of 'yes' and 'no' events. We further recall that we used the conditional l/L-quantiles to predict threshold exceedance by CS.

CS and LCMCK outperformed LUKp and LCKp in predicting exceedance of small (Cd $<$ 1 mg kg^{-1}) and large (Cd $>$ 2 mg kg^{-1}) thresholds, and CS was as good as and LCMCK somewhat worse than the two kriging methods in between (Fig. 7A). Hence, LUKp and LCKp overestimated exceedance for the small and underestimated it for the large thresholds, resulting in a large false alarm rate in the former and a small hit rate in the latter case, deteriorating PSS thereby in both instances. The performance of LUKp and LCKp differed only for the thresholds 2–3 mg Cd kg^{-1} where LCKp was slightly better than LUKp. The PSS of LUK (not shown) did not differ much from that of LUKp, and LCK, due to its positive marginal bias overestimated threshold exceedance over the whole range of cutoffs, resulting in a large false alarm for the small and a large hit rate for the large thresholds, but performing on average worst among all the methods.

The Pierce skill scores showed also for Cu (Fig. 7B) some advantage of CS and LCMCK over the other methods, mostly for small (Cu $<$ 300 mg kg^{-1}) and large (Cu $>$ 1000 mg kg^{-1}) thresholds. The performance of LUKp, LCKp, LUK (the latter not shown) hardly differed. LCK showed the same pattern as for Cd (positive marginal bias, resulting in a large false alarm rate for small and a large hit rate for large thresholds). For Zn the PSS curves of CS and LCKp were quite similar (Fig. 7C), with some advantage of CS over LCKp for thresholds in 250–1000 mg kg^{-1}. Moreover, LUKp and LUK (not shown) predicted threshold exceedance nearly as well, except for the thresholds between 1000 and 2000 mg kg^{-1}. LMCK and LCK (the latter not displayed) showed again the pattern typical for positive marginal bias, and both methods were not better than either LUKp, LUK or LCKp.

Apart from PSS, we computed also the bias scores (BS, Fig. 7D–F) as a function of the thresholds to see whether the methods systematically over- or underestimated the exceedance frequency. A bias score close to one signals absence of systematic errors in a binary classification problem. Predicting exceedance of the large thresholds by the conditional quantiles of CS resulted for all the metals in a large number of false positives, and this increased the bias score

to 2–4. LCK (not shown) performed for all the metals at least as bad as CS, and LCMCK had large BS for large thresholds of Cu and Zn. The strong positive bias of LCK (and partly also LCMCK) was responsible for the large bias scores: If the predictions are on average too large (positive eBIAS) then we expect BS$>$1, irrespective of the magnitude of the threshold. Such a pattern was also evident for the positively biased LCKp and LCMCK predictions of the Cu and Zn contents: For the small to intermediate thresholds, where the false positives of CS did not matter, LCKp and LCMCK usually had the largest bias score.

5. Discussion

The size of the validation data set was rather small in our study. This must be born in mind when we now compare the results of the empirical validation with those of the simulation study by Hofer and Papritz (2010) and make an attempt to distill some general conclusion from the empirical validation.

First, |eBIAS| and eMSPE were among the seven methods generally smallest for CS, and CS predicted threshold exceedance with least conditional bias. This suggests that the Gaussian model was not grossly wrong for the log-transformed metal measurements. Second, in contrast to the simulation study where all the methods predicted the block means practically without bias and where we did not use any back-transformation, LCK in particular and the other constrained predictors somewhat less, tended to be positively biased. Whereas the bias was negligible for Cd, it was pretty obvious for Cu and Zn and contributed noticeably to the eMSPEs. The amplification of the differences between constrained and UK predictions of the log-transformed metal content by the back-transformation leads to marginal bias only if the weighted sums of the GLS residuals (term $\mathbf{c}(\mathbf{s}_{1\ldots n},B)'\Sigma^{-1}(\mathbf{Y}-\mathbf{X}\widehat{\beta}_{\mathrm{GLS}})$ in equation (10)) are dominantly positive (or negative). Since most of the non-zero elements of $\mathbf{c}(\mathbf{s}_{1\ldots n},B)'\Sigma^{-1}$ are positive, this happens when most GLS residuals are themselves positive, i.e., if the GLS trend surface 'underfits' the transformed observations in the vicinity of the validation blocks. The back-transformed kriging predictions are then larger than the GLS predictions. Figures 4B–C show that this was indeed the case for Cu and Zn (positive LUKp – GLS differences with only few exceptions for small metal content). For Cd, however, the LUKp – GLS differences were more balanced, which resulted — in combination with the smaller concentration — in less asymmetry in LCKp – LUKp differences when transforming the CK predictions of the log-transformed metal content back to the original scale (cf. Figures 4G vs. H–I). It thus appears that differential success to fit the data in the vicinity of the validation blocks by the regression model was at least partly responsible for the presence or absence of bias and the partly poor precision of constrained or covariance-matching constrained kriging predictions.

Third, unlike some simulation results reported by Cressie (2006), LUK did not outperform LUKp, and LCKp performed far better than LCK. The poor performance of LCK was caused by much larger K values of CK point than CK block predictions of the log-transformed metal content in combination with underfitting of the trend. Averaging the back-transformed point predictions did not moderate the effect of large K values. Using (18) with point CMCK of log-transformed concentrations did neither offer a clear advantage over LCKp. Hence, the assumption that both measurements and block means follow lognormal laws seems defendable for our study. The rather small ratios of the supports of blocks and observations certainly have contributed to the success of the back-transformation (15).

Fourth, the differences in eMSPE qualitatively agree with theory and our previous simulation results. CS, LUK and LUKp were consistently better than the constrained predictors.

The small to moderate values of K (Table 3) for LCKp suggest that the measurements were at least moderately dense around most validation blocks and that there was little need to amplify the GLS residuals to meet the variance constraint of block CK. However, compared with the dense sampling situation in the simulations, the differences in eMSPE between LCKp and LUKp were rather large, in particular for Cu and Zn. A more detailed analysis revealed that the ratios $\text{eMSPE}_{\text{LCKp}}/\text{eMSPE}_{\text{LUKp}}$ were always larger than the ratios $\text{eMSPE}_{\text{CK}}/\text{eMSPE}_{\text{UK}}$, computed for the log-transformed data. Hence, for Cu and Zn, the back-transformation not only introduced some bias, but it also aggravated the loss of precision of LCKp relative to LUKp.

Fifth, the average magnitude of PSS indicates that all the methods predicted threshold exceedance reasonably well. This was confirmed by the areas (AUC) under the Receiver Operating Characteristic curves (e.g. Fawcett, 2006) that we computed in addition to plots of PSS vs. $T_{[l]}$. AUC ranged from 0.78–0.92 (Table 4). Hosmer and Lemeshow (2000) consider such values as good to excellent in a binary classification problem (AUC $= 1$ means perfect and AUC $= 0.5$ random discrimination). The ranking of the methods in predicting threshold exceedance agrees with their performance in the simulations when many measurements were available around the target blocks: CS performed then better than either CK or UK for predicting exceedance of the large thresholds, and UK was nearly as good as CK over the whole range of the thresholds. Furthermore, BS signalled for CS in the simulations at the large thresholds substantial marginal bias (many false positives). Hence, the results of the validation of threshold exceedance support the view that the density of the measurements was quite large around the validation parcels, in particular northeast of the plant where most parcels lay (Fig. 1). Hofer and Papritz (2010, pp. 644–646) showed that UK and CK predictions of nonlinear functionals of block means then do not differ much because the variance of the UK predictor is not much smaller than the variance of the block mean. Under these circumstances, UK (and LUKp) perform quite well when predicting nonlinear functionals of block means, and the gain in precision, that could be obtained by using more sophisticated (and more demanding) methods is likely to remain rather limited.

To see if the density of the support points had some influence on the quality of the predictions, we divided the Cu calibration data set into 5 parts with about 115 observations in each part, predicted the mean Cu content and threshold exceedance for the 53 validation parcels from each part, and averaged the performance measures for the five parts. Since some point or block CK or CMCK predictions did not exist, we had predictions only for 41 validation parcels from all the five parts. The main difference to the results shown in Figures 5–7 was a substantial negative bias of LUK, which increased also its eMSPE in comparison to LUKp and LCKp. Furthermore, the PSS curve of LUK signalled substantial underestimation of exceedance of large thresholds. The performance of the other kriging methods did not change much, when the number of calibration observations was reduced to one fifth of the full calibration data set. Reducing the calibration information even further by splitting the calibration data set into ten parts left us only with 9 validation parcels for which the predictions always existed, and this was too little to draw any conclusions about the performance of the methods for even sparser calibration data.

6. Conclusions

We presented an empirical validation of predictions of topsoil heavy metal content on parcels of land in an area where a metal smelter had severely polluted the soils. Precise prediction of the pollution level of property is essential under such circumstances because it is a prerequisite for correctly assessing the health risks and mitigating them by appropriate measures. Besides

predictions of the mean metal content we therefore assessed the precision of predictions that parcel means exceed regulatory thresholds. All the predictions were computed from measurements with quasi-point support. To our knowledge, our study is the first attempt to validate predictions by conditional simulations, lognormal customary and several forms of constrained block kriging for a local change of support situation with soil data.

On average, the conditional simulations gave the most precise predictions, both of the block means and of threshold exceedance. However, the advantage of this approach was not pronounced: Lognormal universal block kriging — regardless whether based on the assumption of permanence of lognormality or on averaging lognormal point kriging predictions — predicted the block means nearly as well and was not much worse than either form of constrained kriging or conditional simulations when predicting threshold exceedance. All the methods predicted threshold exceedance with good success.

Averaging lognormal constrained or covariance-matching constrained point predictions did not offer a consistent advantage over customary universal or constrained lognormal block kriging computed on the assumption of permanence of lognormality. However, the latter method did not perform as well as in our previous simulation study. Its mean squared error was quite large, and it had no clear advantage over customary lognormal universal block kriging when predicting threshold exceedance. We think that this result does not reflect a general weakness of the method, but was rather due to the combined effect of a local misfit of the trend model and the back-transformation to the original scale of the measurements. However, further work is needed to corroborate or refute this tentative assessment.

The good performance of lognormal block kriging came rather as a surprise because non-linear transforms of the customary block kriging predictor are commonly thought to be quite strongly biased because the kriging predictions are smoother than the target quantity. Constrained or covariance-matching constrained block kriging and conditional simulations should then have a clear advantage over customary block kriging. But we already found in the simulations that the smoothing bias of block kriging does not matter much if soil sampling is sufficiently dense. In this situation, we can expect only limited gains in the precision of predictions by more sophisticated methods. But the simulations also showed that they pay off when sampling is sparse. Under such circumstances, one should prefer — in spite of the result of this empirical validation study — constrained, covariance-matching constrained kriging or conditional simulations for local change of support problems.

Acknowledgements

We gratefully acknowledge the financial support of the Swiss Federal Office for the Environment (FOEN). Noel Cressie and an anonymous reviewer made useful comments on an earlier version of the article which we gratefully acknowledge.

Aldworth, J., Cressie, N., 2003. Prediction of nonlinear spatial functionals. Journal of Statistical Planning and Inference 112, 3–41.

Altfelder, S., Beyer, C., Duijnisveld, W.H.M., Schneider, J., Streck, T., 2002. Distribution of Cd in the vicinity of a metal smelter: Interpolation of soil Cd concentrations with regard to regulative limits. Zeitschrift für Pflanzenernährung und Bodenkunde 165, 697–705.

Bitterli-Brunner, P., Fischer, H., Herzog, P., 1984. Geologischer Atlas der Schweiz 1:25000 Blatt 1067 Arlesheim. Schweizerische Geologische Kommission.

Bourennane, H., King, D., Couturier, A., Nicoullaud, B., Mary, B., Richard, G., 2007. Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: An empirical comparison of a simulation accounting for single attribute and a simulation accounting for secondary information. Ecological Modelling 205, 323–335.

Brus, D.J., de Gruijter, J.J., Marsman, B.A., Visschers, R., Bregt, A.K., Breeuwsma, A., Bouma, J., 1996. The performance of spatial interpolation methods and choropleth maps to estimate properties at points: A soil survey case study. Environmetrics 7, 1–16.

Buttafuoco, G., Tallarico, A., Falcone, G., 2007. Mapping soil gas radon concentration: A comparative study of geostatistical methods. Environmental Monitoring and Assessment 131, 135–151.

Chan, G., Wood, A.T.A., 1997. An algorithm for simulating stationary gaussian random fields. Journal of the Royal Statistical Society Series C 46, 171–181.

Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Chu, H.J., Lin, Y.P., Jang, C.S., Chang, T.K., 2010. Delineating the hazard zone of multiple soil pollutants by multivariate indicator kriging and conditioned latin hypercube sampling. Geoderma 158, 242–251.

Cressie, N., 1993a. Aggregation in geostatistical problems, in: Soares, A. (Ed.), Geostatistics Tróia '92, Kluwer Academic Publishers, Dordrecht. pp. 25–36.

Cressie, N., 1993b. Statistics for Spatial Data. Wiley, New York. revised edition.

Cressie, N., 2006. Block kriging for lognormal spatial processes. Mathematical Geology 38, 413–443.

Dowd, P.A., 1984. The variogram and kriging: Robust and resistant estimators, in: Verly, G., David, M., Journel, A., Maréchal, A. (Eds.), Geostatistics for Natural Resources Characterization, Dordrecht. D. Reidel Publishing Company. pp. 91–106.

Ersoy, A., Yunsel, T.Y., Atici, Ü., 2008. Geostatistical conditional simulation for the assessment of contaminated land by abandoned heavy metal mining. Environmental Toxicology 23, 96–109.

Faraway, J., 2004. Linear Models with R. Texts in statistical science, Chapman & Hall/CRC, Boca Raton.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Gotway, C.A., Young, L., 2002. Combining incompatible spatial data. Journal of the American Statistical Association 97, 632–648.

Grünenfelder, B., Schmidli, F., 1998. Schwermetallbilanzierung belasteter Böden und ihre Anwendung im Rahmen der sanften Bodensanierung. Diplomarbeit. Eidgenössische Technische Hochschule ETH Zürich, Abteilung für Kulturtechnik und Vermessung.

Hani, A., Pazira, E., 2011. Heavy metals assessment and identification of their sources in agricultural soils of southern tehran, iran. Environmental Monitoring and Assessment 176, 677–691. 10.1007/s10661-010-1671-5.

Hofer, C., Papritz, A., 2010. Predicting threshold exceedance by local block means in soil pollution surveys. Mathematical Geosciences 42, 631–656. Doi:10.1007/s11004-010-9287-4.

Hofer, C., Papritz, A., 2011. constrainedkriging: An r-package for customary, constrained and covariance-matching constrained point or block kriging. Computers & Geosciences In Press, Corrected Proof, –.

Hosmer, D., Lemeshow, S., 2000. Applied Logistic Regression. Wiley New York. 2 edition.

Karanlık, S., Ağca, N., Yalçin, M., 2011. Spatial distribution of heavy metals content in soils of amik plain (hatay, turkey). Environmental Monitoring and Assessment 173, 181–191. 10.1007/s10661-010-1380-0.

Kayser, A., Presler, J., Meuli, R., Kägi, J., 2006. Bodenbelastungsgebiet Dornach - Zusatzuntersuchungen (Projekt P3) - Synthesebreicht. Technical Report. Amt für Umwelt des Kantons Solothurn. http://www.so.ch/fileadmin/internet/bjd/bumaa/pdf/boden/synthesebericht_p3_a.pdf.

Laslett, G.M., 1994. Kriging and splines: An empirical comparison of their predictive performance in some applications (with discussions). Journal of the American Statistical Association 89, 391–400.

Lin, Y.P., Chang, T.K., Shih, C.W., Tseng, C.H., 2002. Factorial and indicator kriging methods using a geographic information system to delineate spatial variation and pollution sources of soil heavy metals. Environmental Geology 42, 900–909.

Lin, Y.P., Chang, T.K., Teng, T.P., 2001. Characterization of soil lead by comparing sequential Gaussian simulation, simulated annealing simulation and kriging methods. Environmental Geology 41, 189–199.

Marchant, B.P., Tye, A.M., Rawlins, B.G., 2011. The assessment of point-source and diffuse soil metal pollution using robust geostatistical methods: a case study in Swansea (Wales, UK). European Journal of Soil Science 62, 346–358.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. *Robust Statistics Theory and Methods*. John Wiley & Sons, Chichester.

McGrath, D., Zhang, C., Owen, T.C., 2004. Geostatistical analyses and hazard assessment on soil lead in silvermines

area, Ireland. Environmental Pollution 127, (239–248).
Moyeed, R.A., Papritz, A., 2002. An empirical comparison of kriging methods for nonlinear spatial point prediction. Mathematical Geology 34, 365–386.
OIS, 1998. Ordinance relating to the impacts on the soils of 1 july 1998. SR 814.12, Swiss Federal Council, Berne, Switzerland. http://www.admin.ch/ch/d/sr/8/814.12.de.pdf.
Papritz, A., Dubois, J.P., 1999. Mapping heavy metals in soil by (non-)linear kriging: An empirical validation, in: Gómez-Hernández, J., Soares, A., Froidevaux, R. (Eds.), geoENV II: Geostatistics for Environmental Applications, Kluwer, Dordrecht. pp. 429–440.
Papritz, A., Herzig, C., Borer, F., Bono, R., 2005. Modelling the spatial distribution of copper in the soils around a metal smelter in northwestern Switzerland, in: Renard, P., Demougeot-Renard, H., Froidevaux, R. (Eds.), Geostatistics for Environmental Applications, Springer-Verlag, Berlin Heidelberg. pp. 343–354.
Paul, R., Cressie, N., 2011. Lognormal block kriging for contaminated soil. European Journal Soil Science 62, 337–345.
Peirce, C.S., 1884. The numerical measure of the success of predictions. Science 4, 453–454.
R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.
Rambeau, C., Föllmi, K., Adatte, T., Matera, V., Steinmann, P., Veuve, P., 2002. Cadmium anomalies in Oolitic carbonates of Bajocian and Oxfordian/Kimmeridgian age in the Swiss and French Jura mountains. Geochimica et Cosmochimica Acta 66, A623. Supplement August 2002.
Rawlins, B.G., Lark, R.M., Webster, R., O'Donnell, K.E., 2006. The use of soil survey data to determine the magnitude and extent of historic metal deposition related to atmospheric smelter emissions across Humberside, UK. Environmental Pollution 143, 416–426.
Rousseeuw, P. an Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Mächler, M., 2011. robustbase: Basic Robust Statistics. R package version 0.7-3.
Saito, H., Goovaerts, P., 2001. Accounting for source location and transport direction into geostatistical prediction of contaminants. Environmental Science & Technology 35, 4823–4829.
Schlather, M., 2001. Simulation of stationary and isotropic random fields. R-news 1, 18–20.
Schlather, M., 2011. RandomFields: Simulation and Analysis of Random Fields. R package version 2.0.45.
Schuhmaker, L., 1981. *Spline Functions: Basic Theory*. John Wiley & Sons, New York.
Schulin, R., Gupta, S.K., 2002. Sanfte Sanierung von schwermetallbelasteten Böden—der Fall Dornach. TerraTech 11, 36–39.
Sollitto, D., Romic, M., Castrignanò, A., Romic, D., Bakic, H., 2010. Assessing heavy metal contamination in soils of the zagreb region (northwest croatia) using multivariate geostatistics. CATENA 80, 182 – 194.
Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. Journal of Soil Science 41, 473–490.
Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press. 3 edition.
Xie, Y., Chen, T., Lei, M., Yang, J., Guo, Q., Song, B., Zhou, X., 2011. Spatial distribution of soil heavy metal pollution estimated by different interpolation methods: Accuracy and uncertainty analysis. Chemosphere 82, 468 – 476.

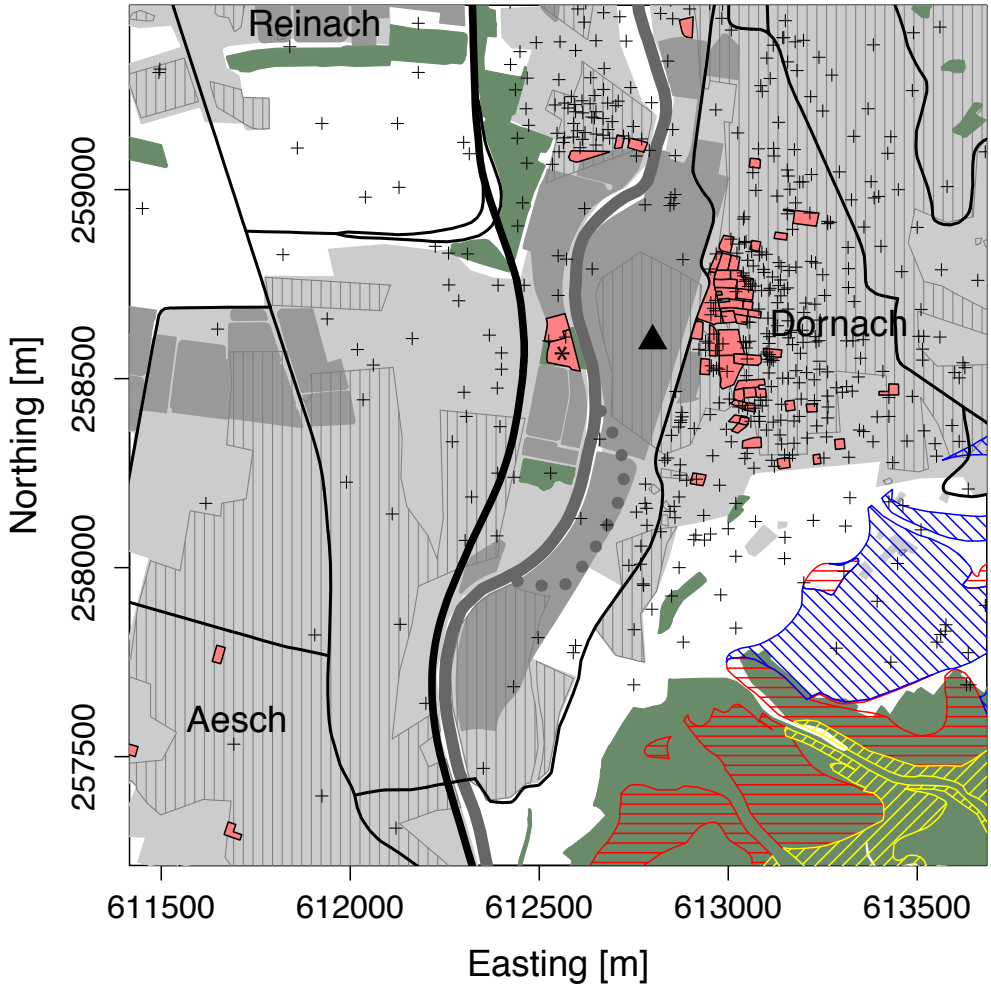

Figure 1: Map of area around metal smelter in Dornach. Sampling sites of the calibration data set are marked by +, validation parcels are shown in light red, the triangle marks the position of the main stack of the plant. Residential areas are shown in light and industrial premises in medium grey, forests by green and agricultural land by white areas (land use in year 2000). Areas already built-up in 1959 are vertically hatched. Areas with Cd-rich parent material are hatched by coloured lines (Lower/Middle Oxfordian [yellow], Bajocian/Upper Oxfordian [red], Lower Rupelian [blue]). The solid lines are the major roads (black) and a river (thick, dark grey). A previous river bed is shown by the thick dotted line.

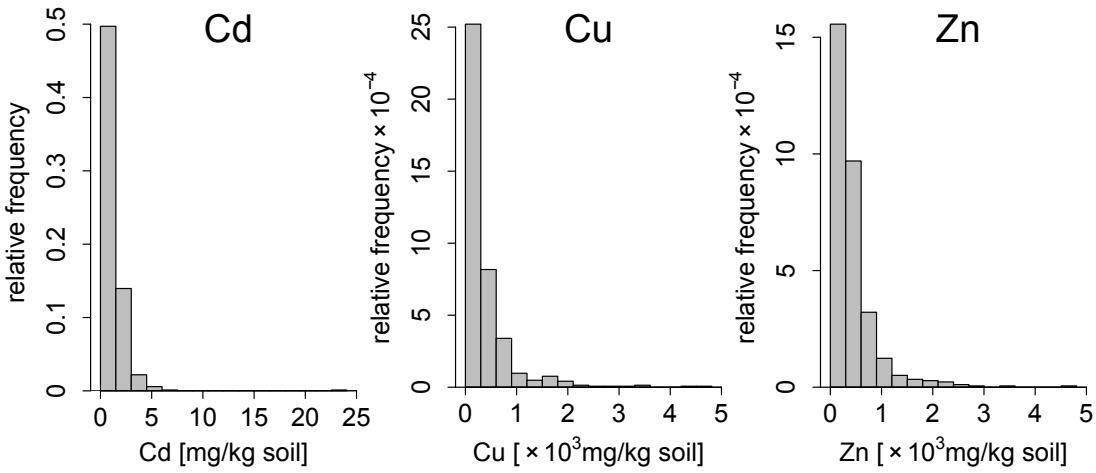

Figure 2: Histograms of the calibration data.

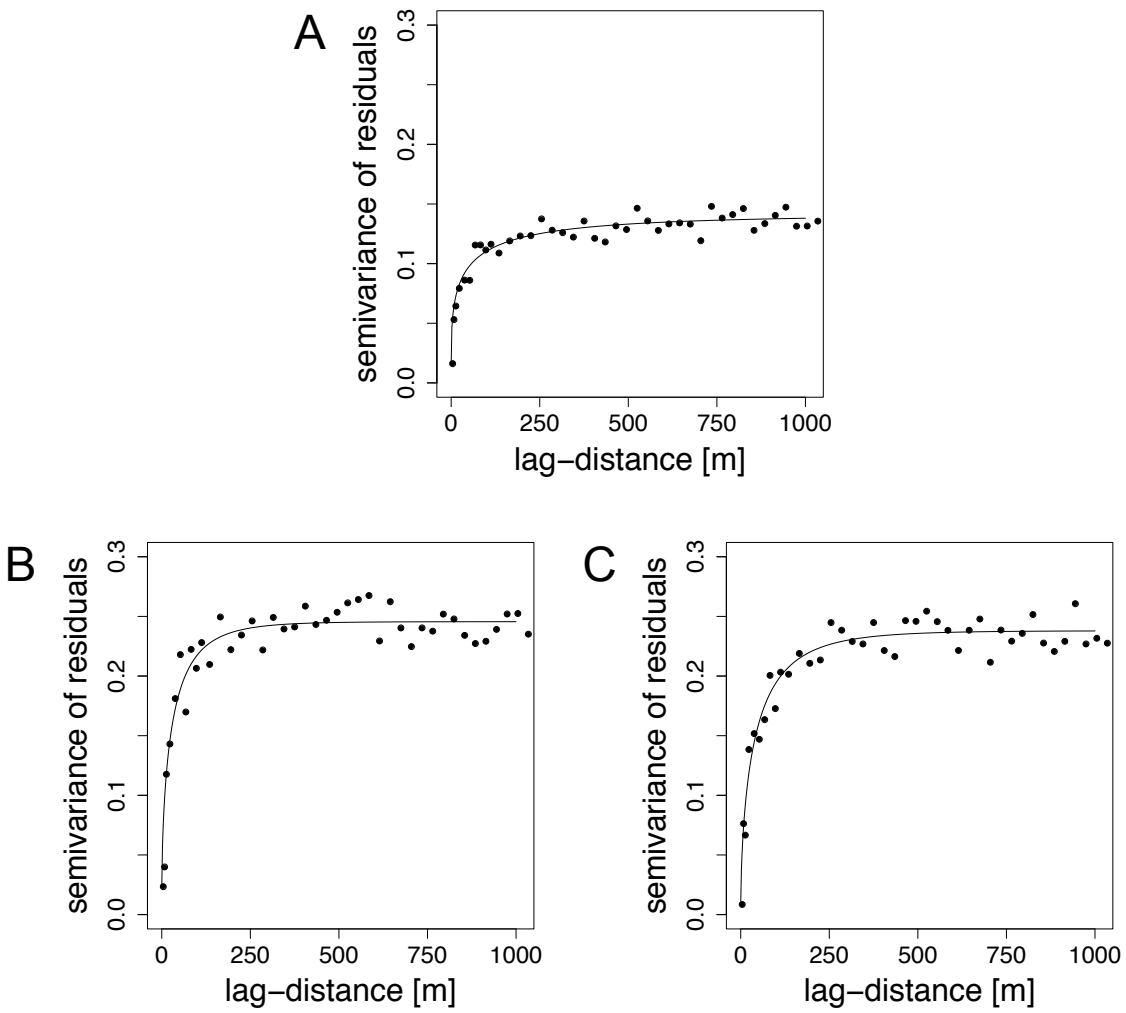

Figure 3: Sample variogram of residuals of robust regression model for A) $\log(\text{Cd}(\mathbf{s}))$, B) $\log(\text{Cu}(\mathbf{s}))$, C) $\log(\text{Zn}(\mathbf{s}))$. The curves show the stable variogram models fitted to the sample variograms.

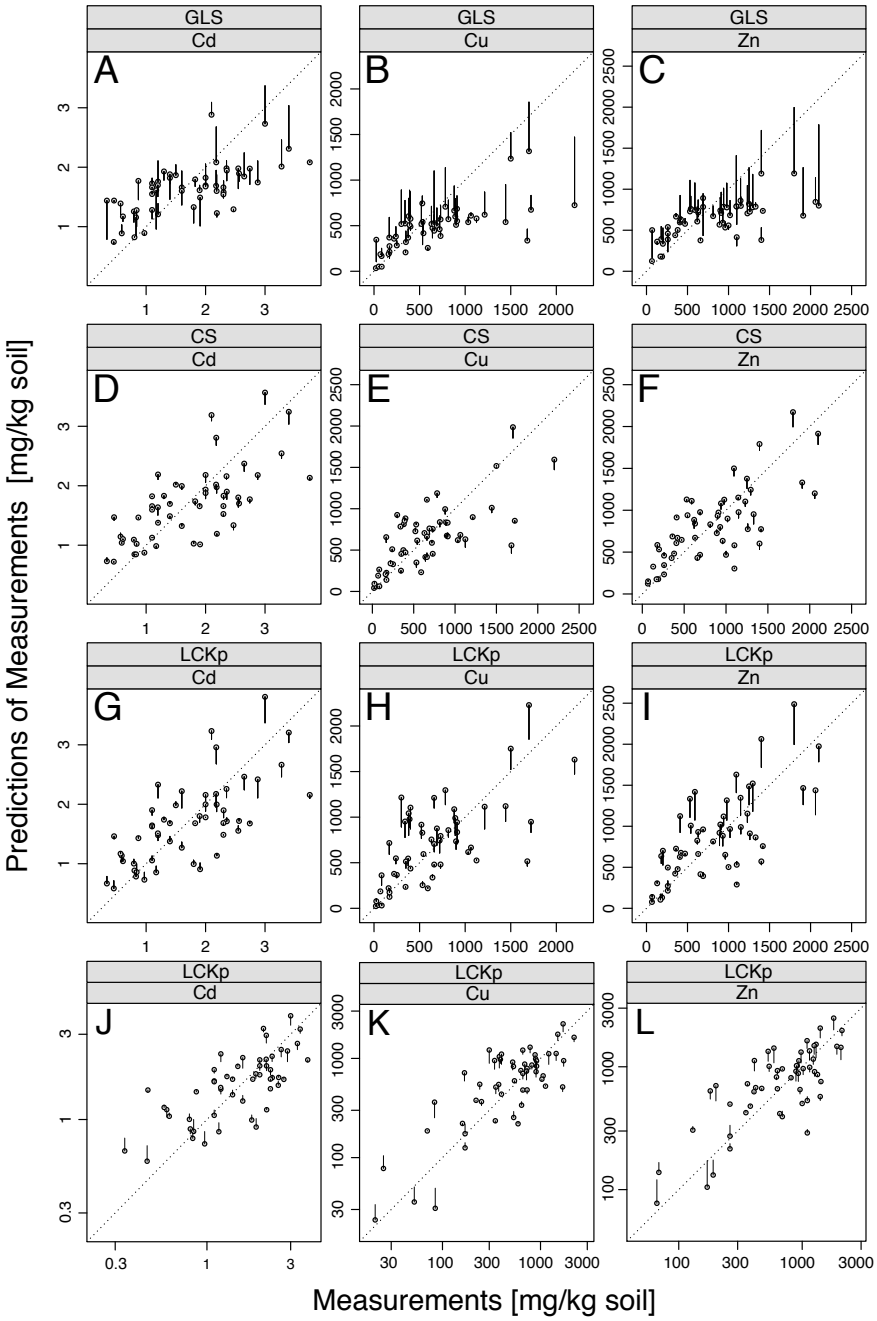

Figure 4: Scatterplots of GLS, CS, LCKp predictions of Cd (A, D, G, J), Cu (B, E, H, K) and Zn (C, F, I, L) contents, plotted against the measured content of the respective metal. The LCKp predictions are either plotted on a linear scale (G–I) or log-scale (J–L). The line segments represent differences in predictions by the respective method and LUKp.

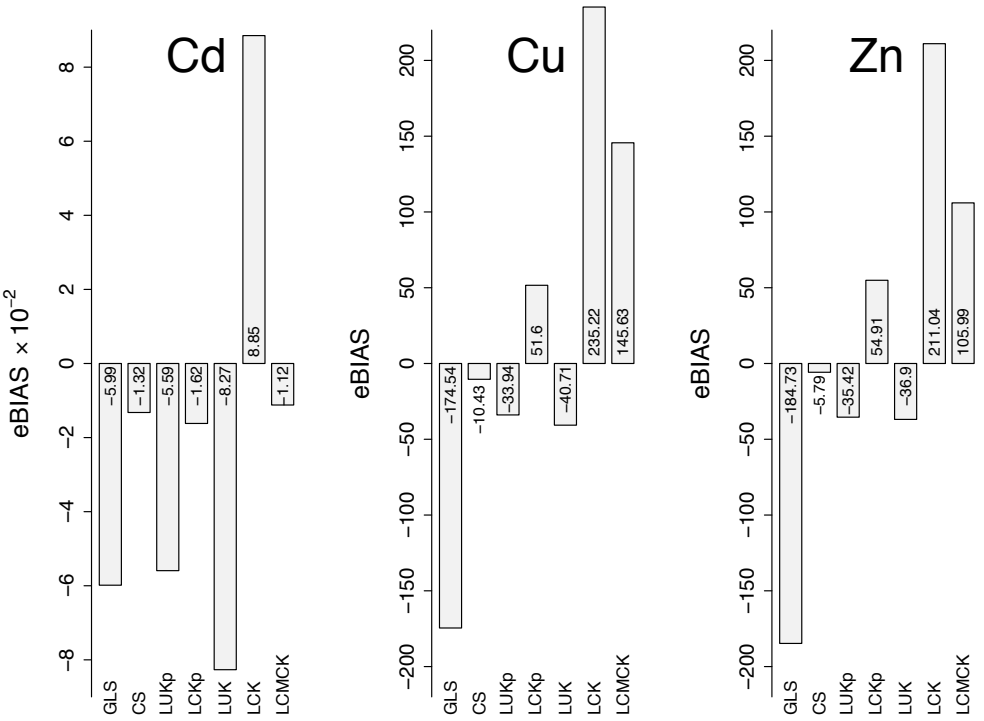

Figure 5: Empirical bias (eBIAS) of conditional simulations (CS), lognormal universal kriging (LUKp, LUK), lognormal constrained kriging (LCKp, LCK), lognormal covariance-matching constrained kriging (LCMCK) and GLS trend surface when predicting the mean topsoil Cd, Cu and Zn contents of 53 validation parcels.

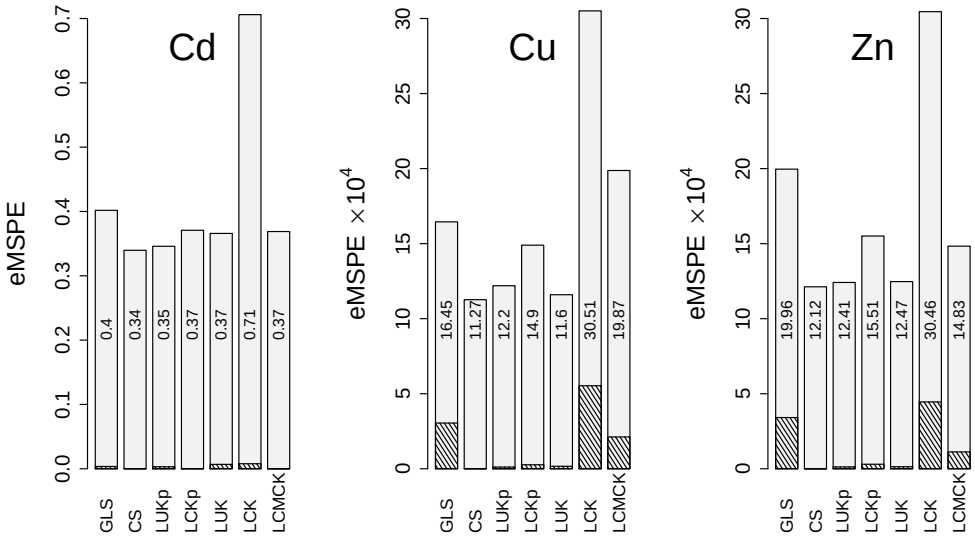

Figure 6: Empirical mean squared prediction error (eMSPE) of conditional simulations (CS), lognormal universal kriging (LUKp, LUK), lognormal constrained kriging (LCKp, LCK), lognormal covariance-matching constrained kriging (LCMCK) and GLS trend surface when predicting the mean topsoil Cd, Cu and Zn contents of 53 validation parcels. The shaded part of the bars shows the contribution of eBIAS^2 to eMSPE.

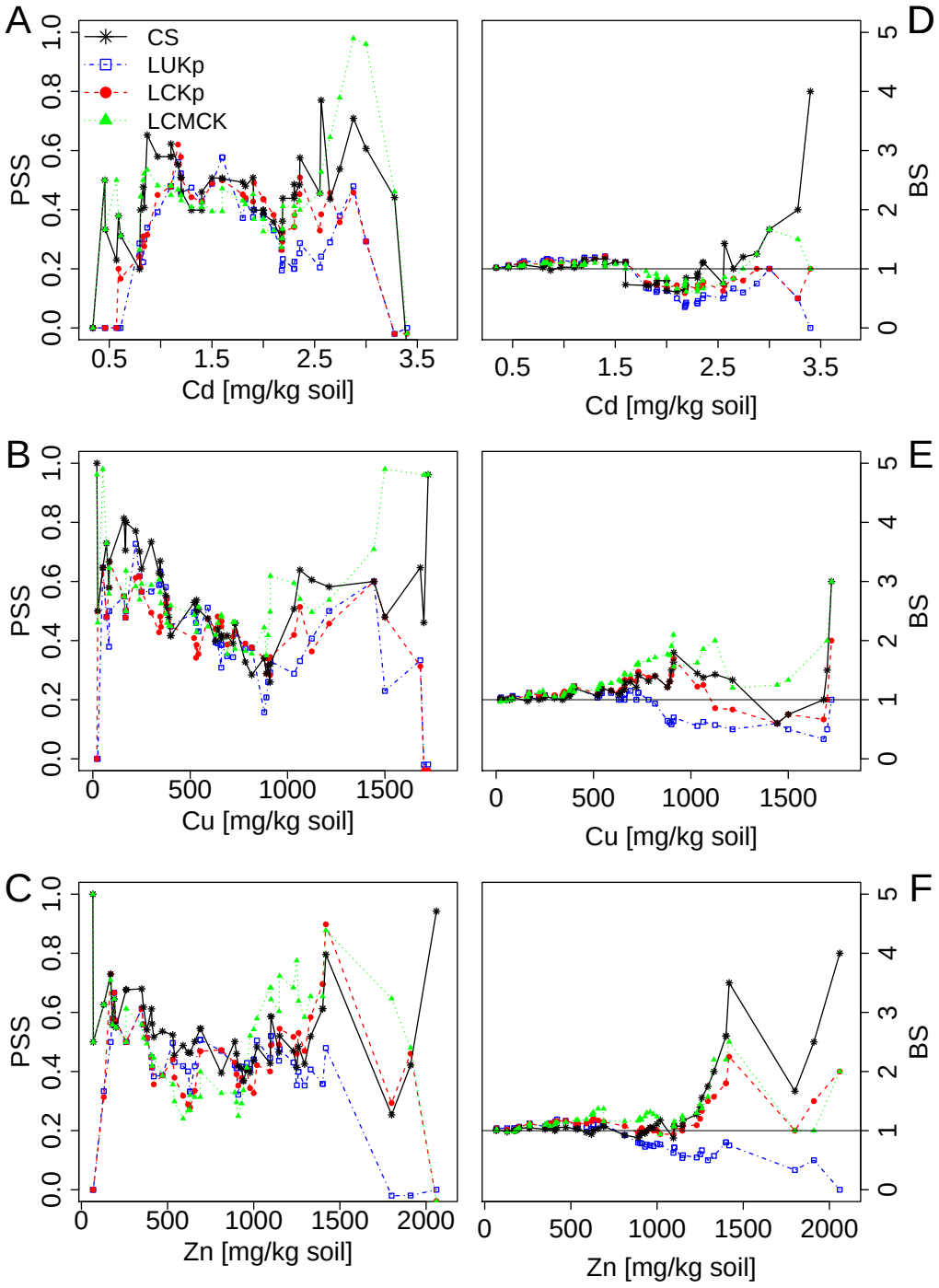

Figure 7: Peirce skill score (PSS) of A) Cd, B) Cu and C) Zn and bias score (BS) of D) Cd, E) Cu and F) Zn vs. the thresholds $T_{[l]}$ for selected prediction methods.

Table 1: GLS estimates of the regression coefficients along with standard errors (SE) for log(Cd), log(Cu), and log(Zn). The coefficients listed for the various levels of the factors are differences between the intercepts for the respective level and the intercept of the reference level 'other bedrock' for geology and 'agriculture/forest' for land use.

	Cd		Cu		Zn	
Covariate and level of factor	$\widehat{\beta}_{\text{GLS}}$	SE	$\widehat{\beta}_{\text{GLS}}$	SE	$\widehat{\beta}_{\text{GLS}}$	SE
$f_0(\mathbf{s})$ N	6.985	1.552	18.519	1.774	16.168	1.944
$f_1(\mathbf{s})$ NE	3.824	1.071	15.298	1.244	12.635	1.379
$f_2(\mathbf{s})$ E	3.670	0.827	11.669	0.868	8.976	0.953
$f_3(\mathbf{s})$ SE	3.508	1.156	15.105	1.216	11.203	1.319
$f_4(\mathbf{s})$ S	4.526	1.700	15.291	2.151	12.622	2.314
$f_5(\mathbf{s})$ SW	-3.990	1.525	7.519	1.783	5.639	1.872
$f_6(\mathbf{s})$ W	4.127	1.189	15.024	1.405	11.198	1.563
$f_7(\mathbf{s})$ NW	0.134	1.320	13.701	1.567	8.042	1.704
$f_0(\mathbf{s})\log(d(\mathbf{s}))$ N	-1.042	0.229	-1.959	0.262	-1.522	0.286
$f_1(\mathbf{s})\log(d(\mathbf{s}))$ NE	-0.613	0.158	-1.625	0.184	-1.129	0.203
$f_2(\mathbf{s})\log(d(\mathbf{s}))$ E	-0.559	0.122	-1.024	0.124	-0.490	0.136
$f_3(\mathbf{s})\log(d(\mathbf{s}))$ SE	-0.512	0.176	-1.533	0.180	-0.839	0.194
$f_4(\mathbf{s})\log(d(\mathbf{s}))$ S	-0.694	0.266	-1.576	0.336	-1.075	0.360
$f_5(\mathbf{s})\log(d(\mathbf{s}))$ SW	0.513	0.218	-0.504	0.257	-0.097	0.268
$f_6(\mathbf{s})\log(d(\mathbf{s}))$ W	-0.641	0.172	-1.623	0.204	-0.926	0.225
$f_7(\mathbf{s})\log(d(\mathbf{s}))$ NW	-0.078	0.195	-1.389	0.233	-0.438	0.252
Land use: built-up before 1960	0.063	0.060	0.087	0.08	0.226	0.078
Land use: built-up after 1960	-0.117	0.060	-0.202	0.08	-0.102	0.079
Indicator for biased Cd measurements	0.338	0.059	–	–	–	–
Geology: Lower/Middle Oxfordian	1.350	0.186	–	–	–	–
Geology: Bajocian/Upper Oxfordian	0.527	0.161	–	–	–	–
Geology: Lower Rupelian	0.237	0.132	–	–	–	–

Table 2: Estimated variogram parameters.

	σ_0^2	σ_1^2	α	κ
log(Cd)	0.016	0.125	44.767	0.415
log(Cu)	0.0234	0.222	32.243	0.643
log(Zn)	0.009	0.229	44.976	0.638

Table 3: Quantiles of the factor K (LCKp, cf. equation 11) for the 53 (Cd: 52) validation parcels.

	probability				
	10 %	25 %	50 %	75 %	90 %
Cd	1.16	1.30	1.44	1.58	1.65
Cu	1.15	1.29	1.50	1.64	1.85
Zn	1.10	1.21	1.37	1.49	1.65

Table 4: Areas under the Receiver Operating Characteristic curves.

	CS	LUKp	LCKp	LUK	LCK	LCMCK
Cd	0.85	0.84	0.84	0.84	0.78	0.83
Cu	0.90	0.84	0.83	0.84	0.88	0.92
Zn	0.88	0.85	0.84	0.85	0.87	0.83