ELSEVIER

Geoderma 102 (2001) 75–100

GEODERMA

www.elsevier.nl/locate/geoderma

Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold

M. Van Meirvenne [a,*], P. Goovaerts [b]

[a] *Department of Soil Management and Soil Care, Ghent University, Coupure 653, 9000 Ghent, Belgium*

[b] *Department of Civil and Environmental Engineering, The University of Michigan, EWRE Building, Ann Arbor, MI 48109-2125, USA*

Received 5 April 2000; received in revised form 18 September 2000; accepted 20 October 2000

Abstract

A non-parametric approach for assessing the probability that heavy metal concentrations in soil exceed a location-specific environmental threshold is presented. The methodology is illustrated for an airborne Cd-contaminated area in Belgium. Non-stationary simple indicator kriging, using a soft indicator coding to account for analytical uncertainty, was used in combination with declustering weights to construct the local conditional cumulative distribution function (ccdf) of Cd. The regulatory Cd contamination threshold (CT) depends on soil organic matter and clay content, which entails that its value is not constant across the study area and also is uncertain. Therefore, soft indicator kriging was used to construct the ccdfs of organic matter and clay. Latin hypercube sampling of the ccdfs of Cd, soil organic matter and clay yielded a map of the probability that Cd concentrations exceed the site-specific CT. Cross-validation showed that the ccdfs provide accurate models of the uncertainty about these variables. At a probability level of 80% we found that the CT was exceeded at 27.3% of the interpolated locations, covering 3192 ha of the study area, illustrating the extent of the pollution. Additionally, a new methodology is proposed to sample preferentially the locations where the uncertainty about the probability of exceeding the CT, instead of the uncertainty about the pollutant itself, is at a maximum. This methodology was applied in a two-stage sampling campaign to identify locations where additional Cd samples should be collected in order to improve the classification into safe and contaminated locations.

Keywords: Cadmium; Contamination threshold; Indicator kriging; Probability maps; Soft indicator coding; Sampling strategy

* Corresponding author. Tel.: +32-9-2646-056; fax: +32-9-2646-247.
E-mail address: marc.vanmeirvenne@rug.ac.be (M. Van Meirvenne).

0016-7061/01/$ - see front matter
PII: S0016-7061(00)00105-1

1. Introduction

There is an increasing awareness that an estimate is of little value in the absence of a measure of the associated uncertainty. This is specially the case of prediction of environmental variables where the prediction uncertainty is required to support decision-making about further management. Over the last 20 years, geostatistical methods, like kriging, have been used successfully to investigate the spatial variability of continuously varying environmental variables and to incorporate this information into mapping (Burrough and McDonnell, 1998). However, the kriging variance has often been misused as a measure of reliability of the kriging estimate. The main limitation of the kriging variance is that, when it is used to calculate the probability of exceedence, it relies on the assumptions of normality of the distribution of prediction errors (as, e.g. in Tiktak et al., 1999) and of homoscedasticity (i.e. the variance of the errors is independent from the data values). These conditions are rarely met for environmental attributes, which typically display highly skewed histograms. An alternative is to use indicator kriging (Journel, 1983), to derive, at each unsampled location, the conditional cumulative distribution function (ccdf) which models the uncertainty about the unknown value. This approach does not rely on an analytical (parametric) modelling of the shape of the error histogram, hence it is referred to as "non-parametric". Furthermore, it can account for measurement errors through a soft indicator coding of observations (Journel, 1986), which contrasts with most studies on heavy metals where the measurement errors were assumed to be negligible (e.g. Goovaerts et al., 1997; Juang and Lee, 1998). Also, the ccdfs can be used to analyse how the uncertainty propagates when several variables are combined (Heuvelink, 1998). This uncertainty propagation can be conducted numerically by sampling the ccdfs of these variables many times to consider all possible combinations.

Uncertainty assessment is not a goal per se, but it is a preliminary step in the decision-making process, such as delineation of hazardous areas. In the process of site characterization and remediation, multistage, or phased, sampling is often conducted so as to validate the result of prior sampling or to improve the cost-effectiveness of a sampling campaign (Englund and Heravi, 1994). Phased sampling involves an interruption of the sampling process until the data are available for estimating contaminant concentrations at unsampled locations, which will guide the selection of locations where additional data are needed. Chien (1998) found that two-stage sampling led to a smaller proportion of locations that were wrongly classified. Different criteria can be used to locate these additional samples. A common approach consists of designing a sampling scheme that minimizes the kriging variance (Webster and Burgess, 1984; Van Groenigen and Stein, 1998). This approach is very convenient for multistage sampling because, as long as the variogram model is unchanged, the impact of sampling on the kriging variance can be assessed a priori (Burgess et al., 1981;

Englund and Heravi, 1994). In this paper we present an alternative approach that is based on the analysis of the ccdfs, and so it is better suited to the presence of heteroscedasticity (i.e. the variance of the estimation errors depends on the actual data values). In particular, a new criterion is introduced to sample preferentially the locations where the uncertainty about the exceedence of the sanitation threshold, instead of the uncertainty about the Cd concentration itself, is at a maximum. In that way, the sampling scheme is tailored for the specific objective of improving the remediation decision instead of improving the accuracy of the prediction itself.

Our research deals with a 216-km^2 study area in Belgium, which was contaminated by airborne cadmium for over a century. The origins of the Cd were three zinc factories. Cd was released in the atmosphere until the 1950s and since the 1970s, the emission reduced drastically due to the use of a hydrothermal extraction process. During the 1980s, under the aegis of the Flemish Executive, the topsoil of more than 1500 vegetable gardens was sampled by the Study Centre for Ecology and Forestry (LISEC), and Cd, together with several other soil properties, was determined. Vegetable gardens were targeted since the direct exposure of human to soil contaminated by Cd is most risky when vegetables (especially leafy crops) grown on such soils are consumed (Chaney, 1990). More recently, a new environmental threshold has been applied (Vlaamse Gemeenschap, 1996) to evaluate the contamination of soils by heavy metals. This threshold was defined as a function of soil organic matter and clay content. So the uncertainty of these soil properties must be incorporated in the evaluation of a contamination by heavy metals as well.

The aim of this paper is to present a non-parametric methodology to assess and combine the uncertainty arising from measurement errors and several spatial predictions into the mapping of the probability that a sanitation threshold is exceeded. Additionally, we will discuss how the results of this uncertainty analysis can be used in the design of two sampling strategies to improve decision-making through the collection of additional data.

2. Materials and methods

2.1. Theory

2.1.1. Modelling of local uncertainty

Consider the problem of modelling the uncertainty about the value of a soil attribute z at the unsampled location $\mathbf{x}_0$ (representing a coordinates' vector). The information available consists of a set of n observations $\{z(\mathbf{x}_\alpha), \alpha = 1,2,\ldots,n\}$ which is considered as a realisation of one set of n spatially correlated random variables $Z(\mathbf{x}_\alpha)$. The uncertainty about the z value at $\mathbf{x}_0$ can

be modelled through a random variable $Z(\mathbf{x}_0)$ that is characterised by its distribution function (Goovaerts, 1997):

$$F(\mathbf{x}_0; z|(n)) = \text{Prob}\{Z(\mathbf{x}_0) \leq z|(n)\}. \quad (1)$$

The function $F(\mathbf{x}_0; z|(n))$ is referred to as a conditional cumulative distribution function, where the notation $|(n)$ expresses the conditioning to the n data $z(\mathbf{x}_\alpha)$. The ccdf fully models the uncertainty at $\mathbf{x}_0$ since it gives the probability that the unknown is no greater than any given threshold z.

Determination of a ccdf is straightforward if an analytical model defined by a few parameters can be adopted. For example, under the multi-Gaussian model, the ccdf is Gaussian (Journel and Huijbregts, 1978, p. 566) with the simple kriging estimate and variance as the mean and variance at this location. A non-parametric approach does not assume any particular shape or analytical expression for $F(\mathbf{x}_0; z|(n))$, hence it does not require the adoption of particular models for the random function and is more flexible. It consists of estimating the value of the ccdf for a series of K threshold values z_k, discretizing the range of variation of z:

$$F(\mathbf{x}_0; z_k|(n)) = \text{Prob}\{Z(\mathbf{x}_0) \leq z_k|(n)\}, \qquad k = 1,2,\ldots,K. \quad (2)$$

The resolution of the discrete ccdf is then increased by interpolation within each class $]z_k,\ z_{k+1}]$ and extrapolation beyond the two extreme threshold values z_1 and z_k.

A non-parametric estimation of ccdf values is based on the interpretation that the conditional expectation of an indicator random variable $I(\mathbf{x}_0; z_k)$ given the information (n):

$$F(\mathbf{x}_0; z_k|(n)) = E\{I(\mathbf{x}_0; z_k)|(n)\} \quad (3)$$

with $I(\mathbf{x}_0; z_k) = 1$ if $Z(\mathbf{x}_0) \leq z_k$ and zero otherwise, can be considered as the conditional probability in Eq. (2). Ccdf values can thus be estimated by interpolation of indicator transforms of data, for which we used indicator kriging (Journel, 1983).

2.1.2. Indicator coding

The indicator approach requires a preliminary coding of each observation $z(\mathbf{x}_\alpha)$ into a series of K values indicating whether the threshold z_k is exceeded or not. If the measurement errors are assumed negligible compared to the spatial variability, observations are coded into hard (0 or 1) indicator data:

$$i(\mathbf{x}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{x}_\alpha) \leq z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1,2,\ldots,K. \quad (4)$$

To account for the uncertainty arising from analytical errors, we propose to replace $z(\mathbf{x}_\alpha)$ by a Gaussian distribution centred on $z(\mathbf{x}_\alpha)$ (assuming no bias) and with a standard deviation $s(\mathbf{x}_\alpha) = \text{CV}z(\mathbf{x}_\alpha)$, where CV is the coefficient of

variation of the analytical procedure (repeatability). The indicator coding thus becomes:

$$i(\mathbf{x}_\alpha; z_k) = N\{(z_k - z(\mathbf{x}_\alpha))/s(\mathbf{x}_\alpha)\} \qquad k = 1,2,\ldots,K \tag{5}$$

where $N\{.\}$ is the standard normal cumulative distribution function. Unlike the hard indicator coding (Eq. (4)), coding according to Eq. (5) yields indicators valued between 0 and 1, referred to as soft indicators (Journel, 1986). The difference between a hard and soft indicator coding is illustrated for a clay observation, $z = 3.3$ dag kg^{-1}, determined with an analytical repeatability of 4.7%:

z_k (dag kg^{-1}):	1.6	1.9	2.3	2.6	2.8	3.1	3.5	3.8	4.3
hard $i(z_k)$:	0	0	0	0	0	0	1	1	1
soft $i(z_k)$:	0	0	0	0	0.001	0.099	0.901	0.999	1

The nine thresholds z_k correspond to the nine deciles of the sample distribution of clay used in the subsequent case study.

2.1.3. Indicator kriging

At any unsampled location $\mathbf{x}_0$, each of the K ccdf values can be estimated as a linear combination of indicator transforms of neighbouring observations. The ordinary indicator kriging estimator for threshold z_k is:

$$\left[F\{\mathbf{x}_0; z_k|(n)\}\right]^* = \sum_{\alpha=1}^{n} \lambda_\alpha(z_k) i(\mathbf{x}_\alpha; z_k). \tag{6}$$

The weights $\lambda_\alpha(z_k)$ are obtained by solving the following ordinary indicator kriging system of $(n+1)$ equations (Goovaerts, 1997):

$$\begin{cases} \sum_{\beta=1}^{n} \lambda_\beta(z_k)\gamma_I(\mathbf{x}_\alpha - \mathbf{x}_\beta; z_k) - \psi(z_k) = \gamma_I(\mathbf{x}_\alpha - \mathbf{x}_0; z_k) & \forall \alpha = 1 \text{ to } n \\ \sum_{\beta=1}^{n} \lambda_\beta(z_k) = 1 \end{cases} \tag{7}$$

where $\psi(z_k)$ is a Lagrange parameter. The only information required by the kriging system are K indicator variogram values for different lags, and these are derived from the variogram model $\gamma_I(\mathbf{h}; z_k)$ fitted to experimental values computed as:

$$\gamma_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \{i(\mathbf{x}_\alpha; z_k) - i(\mathbf{x}_\alpha + \mathbf{h}; z_k)\}^2. \tag{8}$$

Because of the impact of wind direction and location of factories on the spatial distribution of Cd, these data display a strong spatial trend that needs to

be taken into account in the interpolation procedure. Consequently, simple indicator kriging with varying local means (Goovaerts and Journel, 1995) was used. Therefore, the estimator of Eq. (6) is re-written as:

$$\left[F\{\mathbf{x}_0;z_k|(n+1)\}\right]^* = y(\mathbf{x}_0;z_k) + \sum_{\alpha=1}^{n} \lambda_\alpha(z_k)\{i(\mathbf{x}_\alpha;z_k) - y(\mathbf{x}_\alpha;z_k)\} \quad (9)$$

where $y(\mathbf{x}_0;z_k)$ is the local mean of the soft indicator for threshold z_k and location $\mathbf{x}_0$ (see the cadmium section for more details about how these local means were derived). The weights $\lambda_\alpha(z_k)$ are obtained by solving a simple indicator kriging system:

$$\sum_{\beta=1}^{n} \lambda_\beta(z_k) C_R(\mathbf{x}_\alpha - \mathbf{x}_\beta;z_k) = C_R(\mathbf{x}_\alpha - \mathbf{x}_0;z_k) \quad \forall \alpha = 1 \text{ to } n \quad (10)$$

where $C_R(\mathbf{h};z_k)$ is the autocovariance of the residual random function $R(\mathbf{x};z_k) = I(\mathbf{x};z_k) - y(\mathbf{x};z_k)$. The residual covariance is typically derived as: $C_R(0) - \gamma_R(\mathbf{h};z_k)$ where the residual variogram model is fitted to experimental values obtained from:

$$\hat{\gamma}_R(\mathbf{h};z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \{r(\mathbf{x}_\alpha;z_k) - r(\mathbf{x}_\alpha + \mathbf{h};z_k)\}^2 \quad (11)$$

with the residuals $r(\mathbf{x}_\alpha;z_k) = i(\mathbf{x}_\alpha;z_k) - y(\mathbf{x}_\alpha;z_k)$.

At each location $\mathbf{x}_0$, the series of K ccdf values must be valued within [0, 1] and be a non-decreasing function of the threshold value z_k, i.e. $[F(\mathbf{x}_0;z_k|(n))]^* \leq [F(\mathbf{x}_0;z_{k'}|(n))]^* \forall z_{k'} > z_k$. These conditions are not necessarily satisfied because kriging weights can be negative and therefore the kriging estimate is a non-convex linear combination of the conditioning data. Following Deutsch and Journel (1998), the first constraint was met by resetting the estimated probabilities outside [0, 1] to the nearest bound, 0 or 1. Order relation deviations between successive ccdf values were corrected using the average of an upward/downward correction. Last, the complete local distribution was retrieved from the set of K ccdf values by linear interpolation between the quantiles as provided by the sample distribution (in case of preferentially clustered observations, declustering weights were taken into account). As discussed in Goovaerts (1999), a limitation of the indicator approach with respect to a multi-Gaussian approach is this a posteriori correction of order relation deviations although these are generally of small magnitude (around 0.01–0.03, Goovaerts, 1994) and should not affect the optimal property of the indicator kriging estimator.

2.1.4. Using local uncertainty models

Knowledge of the ccdf $F(\mathbf{x}_0;z|(n))$ at $\mathbf{x}_0$ allows one to do the following.

(1) Assess the probability of exceeding a critical threshold z_c at $\mathbf{x}_0$:

$$\text{Prob}\{Z(\mathbf{x}_0) > z_c|(n)\} = 1 - F(\mathbf{x}_0;z_c|(n))\} \quad (12)$$

(2) Estimate the unknown value $z(\mathbf{x}_0)$. For example, using a least-squared error criterion amounts at estimating that value by the mean of the ccdf, called E-type estimate (Deutsch and Journel, 1998):

$$z_E^*(\mathbf{x}_0) = \int_{-\infty}^{+\infty} z \mathrm{d}F(\mathbf{x}_0; z|(n)). \tag{13}$$

Similarly, the conditional variance of the ccdf can be calculated.

(3) Generate a series of L simulated values $z^{(l)}(\mathbf{x}_0)$ through a random sampling of the ccdf:

$$z^{(l)}(\mathbf{x}_0) = F^{-1}(\mathbf{x}_0; p^{(l)}|(n)) \qquad l = 1,2,\ldots,L \tag{14}$$

where $p^{(l)}$ are L independent random numbers uniformly distributed in [0, 1]. This procedure is called a Monte-Carlo simulation (Heuvelink, 1998). The set of simulated values can then be used as input to any function $f(.)$, e.g. $Y = f(Z)$, allowing the uncertainty about the output variable Y to be modelled numerically through the distribution of y-values, $y^{(l)}(\mathbf{x}_0) = f(z^{(l)}(\mathbf{x}_0))$. For complex functions $f(.)$ it becomes difficult and time-consuming to generate the large number of simulated values required by a Monte-Carlo analysis of uncertainty propagation. The Latin hypercube sampling (McKay et al., 1979) is a more efficient method of sampling probability distributions (Luxmoore et al., 1991; Pebesma and Heuvelink, 1999). The idea is to divide each ccdf into N equal probability classes which are sampled once to generate a set of N input values. This approach ensures that the ccdf is represented in its entirety in the input to function $f(.)$, and it usually requires a much smaller sample than the traditional Monte Carlo simulation for a given degree of precision.

2.2. Study area, database and sanitation threshold

The study area is located in the Northeast of Belgium (Fig. 1). The Belgian Lambert72 (L72) co-ordinate system was used to geo-reference all samples since it is a metric system facilitating the manipulation of spatial vectors.

Three data sets were used: 1553 analyses of total Cd performed in top soils of vegetable gardens, 1378 soil organic matter measurements collocated with the Cd observations, and 314 top soil clay determinations available from the National Soil Survey database. Cd was determined by atomic absorption spectrometry after extraction by concentrated HNO_3. The repeatability of this procedure was reported to be 7.8% (OFEFP, 1993). Soil organic matter was obtained as $1.724 \times C$, C being measured by the conventional Walkley and Black procedure. We repeated this analysis five times for five samples, leading to an average repeatability of 2.2%. A similar procedure was followed by Van Meirvenne (1991) for the conventional pipette procedure used to determine the

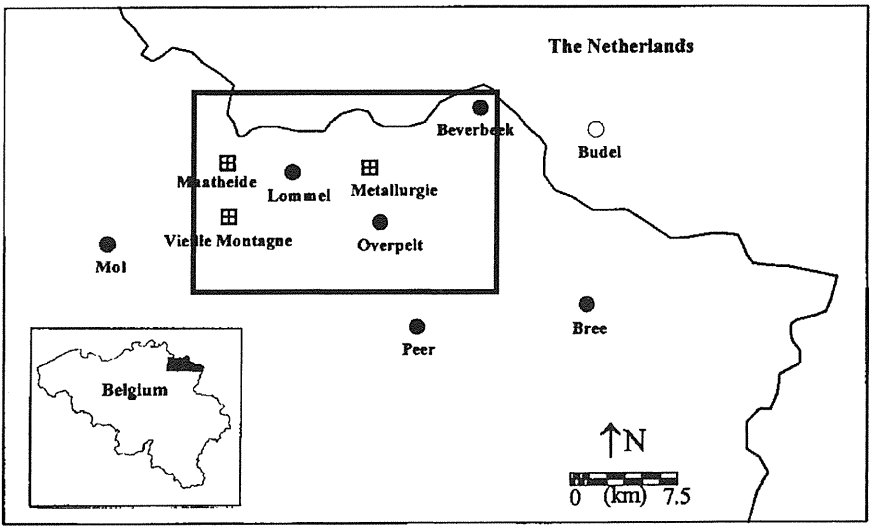

Fig. 1. Location of the study area (rectangle), the three zinc factories (crossed squares) and the major communities (black dots) in Belgium. The coordinates of the lower left corner of the study area are: 5°12′00″E, 51°9′40″N (*X*: 208000 m L72, *Y*: 206000 m L72) and of the top right corner: 5°27′31″E, 51°16′06″N (*X*: 226000 m L72, *Y*: 218000 m L72).

clay content. He found a repeatability of 4.7% for sandy soils. The area contains mainly acid sandy Spodosols with a dominant 100–200 μm sand fraction.

Location maps of the three data sets are given in Fig. 2. The strong spatial clustering of the Cd and soil organic matter data is due to the preferential location of vegetable gardens in communities or along roads. To obtain a histogram and descriptive statistics that are representative for the region, the Cd and SOM data were declustered using square cells of increasing dimension. The goal is to give less weight to redundant (clustered) data located into densely sampled cells (Deutsch and Journel, 1998). Because vegetable gardens with a higher Cd content have been preferentially sampled, declustering leads to a smaller average Cd concentration. The smallest declustered mean Cd content (2.8 mg kg^{-1}) was found for a cell size of 3800 m (the sample mean of the equally weighted distribution was 4.1 mg kg^{-1}). The declustered histogram of Cd (Fig. 2) indicates that the distribution is strongly positively skewed, with extreme values ranging from 0.2 to 70.5 mg kg^{-1}. Notwithstanding the similar spatial configuration, clusters of high or low values were not detected for soil organic matter. Therefore, all observations of soil organic matter were equally weighted and its histogram is also presented in Fig. 2. The soil organic matter distribution is slightly positively skewed, with a mean of 7.0 dag kg^{-1} and values ranging from 1.5 to 22.4 dag kg^{-1}. Soil surveyors of the National Soil Survey collected 314 topsoil samples more or less evenly distributed, and these were analysed for their clay content. The average clay content is 3.0 dag kg^{-1}

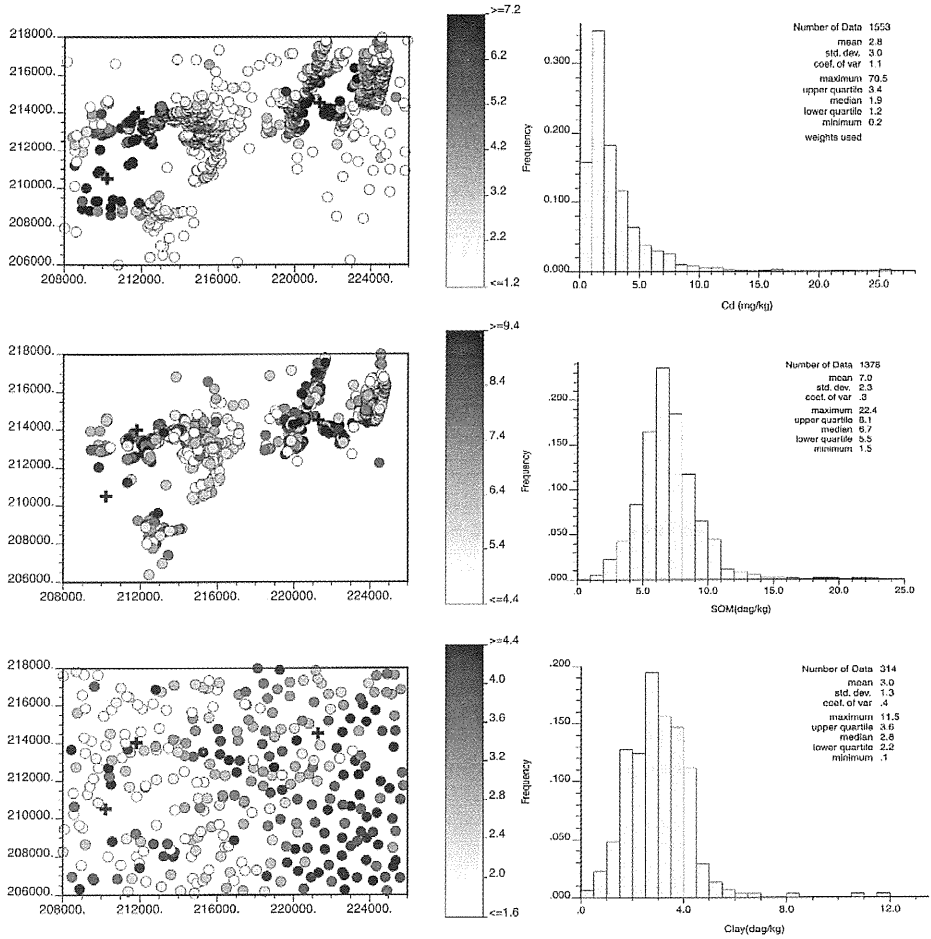

Fig. 2. Location maps (left) and histograms (right) for the Cd (top), soil organic matter (middle) and clay (bottom) data sets. Cd values exceeding the largest value on the axis (25 mg kg^{-1}) were grouped in the last class of the histogram. Location maps were grayscaled between the 10th and 90th percentiles to improve the contrast. Values below the first percentile, or above the last percentile, received the same colour as these percentiles, respectively. Crosses locate the three zinc factories (see Fig. 1).

and the distribution is also slightly positively skewed with extreme values of 0.1 and 11.5 dag kg^{-1}. These samples were taken in non-vegetable gardens and they were also analysed for soil organic matter. Their mean soil organic matter content (4.6 dag kg^{-1}) was lower than the mean soil organic matter of the vegetable gardens, which indicates that the intensive soil use in vegetable gardens increased the soil organic matter content on average by 2.4 dag kg^{-1}.

The Flemish government published a so-called contamination threshold (CT) to evaluate the pollution of soils by heavy metals (Vlaamse Gemeenschap,

1996). It is used to decide whether a soil is suitable for a particular land use (concentration < CT) or whether sanitation measures (like cleaning up) are required (concentration > CT). It is computed according to:

$$\mathrm{CT}(C,O) = N(10,2)\left(\frac{a + bC + cO}{a + 10b + 2c}\right) \tag{15}$$

where C is the clay content (dag kg^{-1}), O the soil organic matter (dag kg^{-1}), CT(C,O) is the CT, and $N(10,2)$, a, b and c are parameters depending on the type of heavy metal and the type of land use. For Cd, $a = 0.4$, $b = 0.03$, $c = 0.05$ and for agricultural land use (including vegetable gardens), $N(10,2) = 2$ mg kg^{-1}.

3. Exceedence of the location-specific sanitation threshold

3.1. Cadmium

3.1.1. Soft indicator coding

The nine deciles of the declustered sample distribution of Cd were used as thresholds z_k. Using Eq. (5), all observations were soft indicator coded with a CV of 7.8%.

3.1.2. Spatial trend—local means

Two processes were considered to be responsible for the clearly observable spatial trend in the Cd data (Fig. 2): the dominant winds and the distance to the sources (the three factories) causing a dilution effect.

A wind rose from a climatic station (Kleine Brogel) near the study area is given in Fig. 3. The dominant winds blow to the north to east directions. This

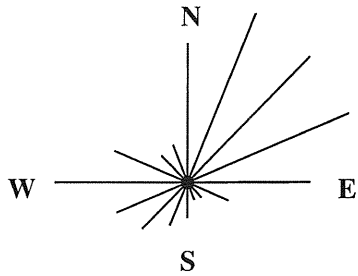

Fig. 3. Wind rose of annual frequencies of wind blowing to a particular direction (longest line—NNE—corresponds to 12.5%).

wind rose was split into four directional classes (4.4% of the days of the year it is windless) defined as:

- High frequency (54.9%): angle interval 348° to 101° (E = 0°)
- Medium frequency (24.4%): angle interval 146° to 236°,
- Relatively low frequency (11.1%): angle interval 236° to 348°, and
- Low frequency (5.2%): angle interval 101° to 146°.

Nine distance classes were considered around each factory: 0–499, 500–999, 1000–1499, 1500–1999, 2000–2999, 3000–3999, 4000–4999, 5000–5999 and ≥ 6000 m. Assuming that the major source of contamination for any sampled

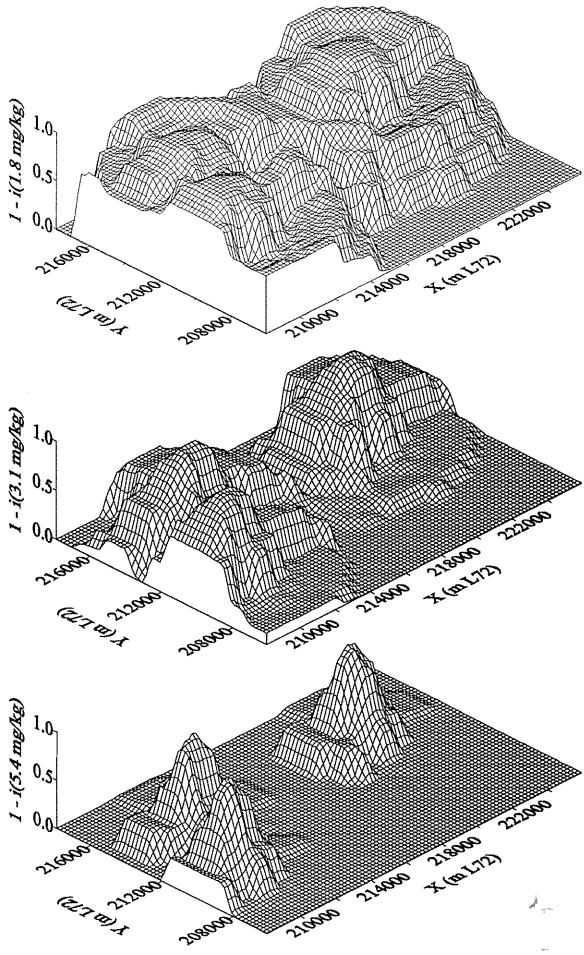

Fig. 4. Trend surfaces (corresponding to 1—the local mean indicator) of the probability to exceed the 20% (1.8 mg kg^{-1}, top), the 50% (3.1 mg kg^{-1}, middle) and the 80% (5.4 mg kg^{-1}, bottom) percentile of the Cd distribution.

location $\mathbf{x}_\alpha$ is the closest of the three factories, each indicator data $i(\mathbf{x}_\alpha; z_k)$ was assigned to one directional and one distance class. For each threshold value z_k, the indicators were averaged within each of these 36 classes resulting in a $9 \times 4 \times 9$ look-up table of local means describing the spatial trend of Cd concentrations. To avoid abrupt changes across class boundaries, these local means were smoothed by interpolation (using ordinary kriging with a linear variogram model). Fig. 4 shows the resulting trend surfaces (plotted as $1 - i(\mathbf{x}; z_k)$) for the 2nd, 5th and 8th decile value z_k. As the threshold value increases, the spatial continuity of the local means decreases dramatically due to the concentration of higher Cd values in the vicinity of the factories and along the most frequent wind directions.

3.1.3. Indicator variograms of residuals

For every z_k the local mean was subtracted from the indicators $i(\mathbf{x}_\alpha; z_k)$ yielding nine sets of residuals $r(\mathbf{x}_\alpha; z_k)$. The omnidirectional variograms of these residuals were calculated (Eq. (11)) and fitted by a combination of a nugget effect and a spherical (first four) or an exponential (last five) model (McBratney and Webster, 1986) (Table 1 and Fig. 5). The gradual increase of the ratio of the nugget effect vs. the sill from 43% to 83% and the corresponding decrease in the range reflect a gradual spatial destructuring as the Cd values increase.

3.1.4. Simple indicator kriging with varying local means and ccdfs

Ccdf values of Cd were estimated by simple indicator kriging with varying local means (Eqs. (9) and (10)). The latter were provided by the trend surfaces at the nodes of a 200 m × 200 m grid. A search radius of 1200 m (corresponding to the largest range of the fitted variograms, see Table 1) was used and a minimum of four neighbouring observations was required before an interpola-

Table 1
Parameters of the nine fitted indicator variograms of the Cd residuals (Fig. 5)

z_k (mg kg^{-1})	Nugget	Model[a]	Sill	Range (m)
1.2	0.042	Sph	0.055	1165
1.8	0.064	Sph	0.074	1220
2.2	0.081	Sph	0.071	1050
2.6	0.089	Sph	0.069	900
3.1	0.077	Exp	0.077	825
3.6	0.081	Exp	0.057	915
4.3	0.074	Exp	0.027	700
5.4	0.052	Exp	0.011	170
7.3	0.025	Exp	0.009	85

[a]Sph: spherical, Exp: exponential.

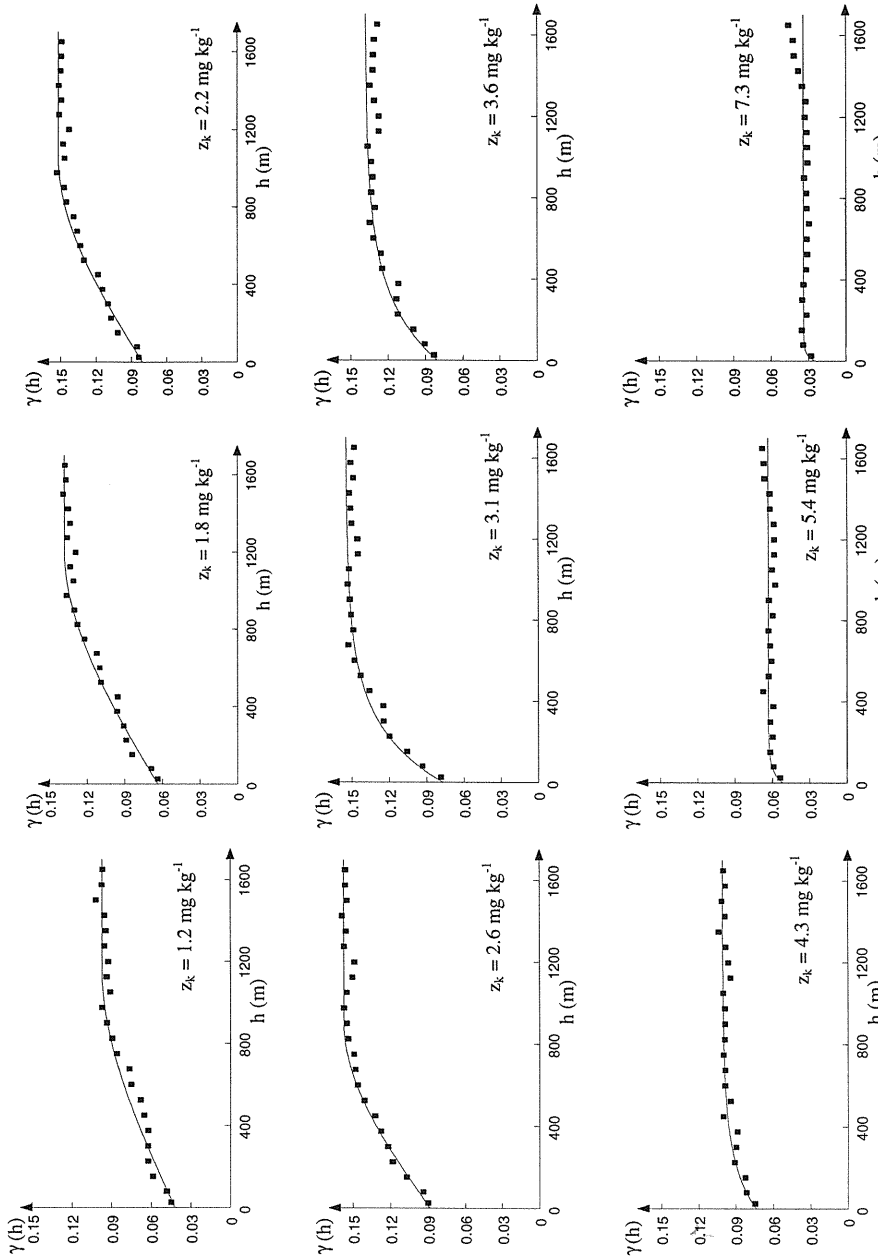

Fig. 5. Indicator variograms of Cd residuals $r(\mathbf{x}_\alpha; z_k)$ for the nine thresholds z_k.

tion was accepted. The number of neighbours was limited to the closest 25 if more were available within the search radius. Due to the uneven spatial coverage of the Cd observations this condition was fulfilled at only 3164 out of 5400 grid nodes. At each of these locations, the interpolated probabilities were used to construct the ccdfs. The means of the ccdfs (Eq. (13)) are mapped in Fig. 6. As expected, the largest Cd concentrations were found near the factories with extensions along the major wind directions.

3.2. Soil organic matter and clay

Due to the absence of a dominant spatial trend for soil organic matter and clay (see Fig. 2), the ccdfs of both variables were obtained by ordinary indicator kriging (Eq. (7)). Again, nine thresholds corresponding to the nine deciles (0.1 to 0.9) of the sample distributions (Fig. 2), were used for the soft indicator coding (Eq. (5)) using the reported repeatabilities. Isotropic indicator variograms (not shown) were computed and modelled for both variables and each of the nine thresholds. Because of the more restricted spatial coverage of soil organic matter compared to Cd, only 2925 grid nodes could be interpolated. At those locations, the ccdfs of soil organic matter and clay were constructed using the same options (to correct order relation problems and to extrapolate at the tails) as for Cd.

3.3. Cross-validation of spatial predictions of Cd, soil organic matter and clay

Before making any decision on the basis of uncertainty models it is critical to evaluate how well the ccdfs capture the uncertainty about the unknown values.

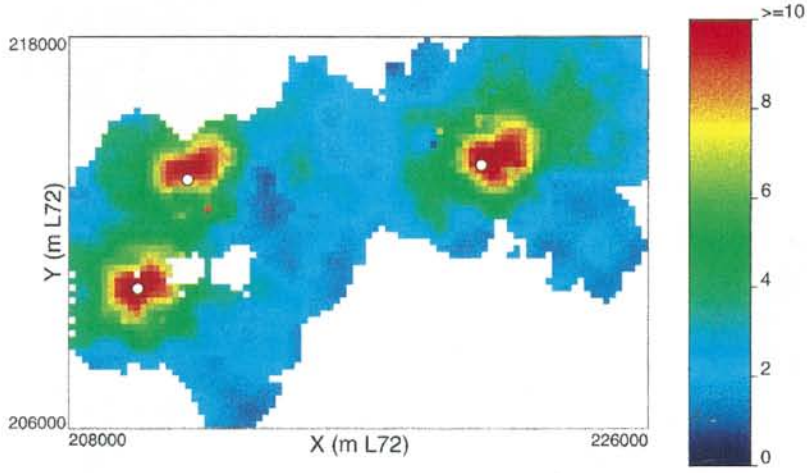

Fig. 6. Map of Cd (mg kg^{-1}) E-type estimates (circles locate factories; values larger than 10 were coloured as 10, largest value = 10.9 mg kg^{-1}; white areas could not be interpolated under the imposed conditions).

As for spatial interpolation, a classical approach is to compare geostatistical predictions with observations that have been temporarily removed one at a time (leave-one-out or cross-validation approach). The major difficulty resides in the selection of performance criteria for ccdfs modeling.

At any test location $\mathbf{x}_0$, a series of symmetric p-probability intervals (PI) can be constructed by identifying the lower and upper bounds to the $(1-p)/2$ and $(1+p)/2$ quantiles of the ccdf $F(\mathbf{x}_0;z|(n))$, respectively. For example, the $0.5-$PI will be bounded by the lower and upper quartiles $[F^{-1}(\mathbf{x}_0;\ 0.25|(n)),\ F^{-1}(\mathbf{x}_0;\ 0.75|(n))]$. A correct modelling of local uncertainty would entail that there is a 0.5 probability that the actual z-value at $\mathbf{x}_0$ falls into that interval or, equivalently, that over the study area 50% of the $0.5-$PI include the true value. If a set of ccdfs have been derived independently from z measurements (e.g. using cross-validation or jackknife) at N data locations $\mathbf{x}_j$, the fraction of true values falling into the symmetric $p-$PI is readily computed as:

$$\xi(p)=\frac{1}{N}\sum_{j=1}^{N}\xi(\mathbf{x}_j;p)\quad\forall\, p\in[0,1] \tag{16}$$

with:

$$\xi(\mathbf{x}_j;p)=\begin{cases}1 & \text{if } z(\mathbf{x}_j)\in\left[F^{-1}\left(\mathbf{x}_j;(1-p)/2|(n)\right),F^{-1}\left(\mathbf{x}_j;(1+p)/2|(n)\right)\right]\\ 0 & \text{otherwise}\end{cases} \tag{17}$$

To account for measurement errors, $\xi(\mathbf{x}_j;p)$ is here computed as:

$$\xi(\mathbf{x}_j;p)=\mathrm{Prob}\left\{F^{-1}\left(\mathbf{x}_j;(1-p)/2|(n)\right)<z(\mathbf{x}_j)\le F^{-1}\left(\mathbf{x}_j;(1+p)/2|(n)\right)\right\}. \tag{18}$$

The closeness of the estimated (experimental) and expected (theoretical) fractions can be assessed using the goodness statistics G (Deutsch, 1997) defined as:

$$G=1-\int_0^1[3a(p)-2][\,\xi(p)-p]\,\mathrm{d}p \tag{19}$$

where $a(p)$ is an indicator function defined as:

$$a(p)=\begin{cases}1 & \text{if } \xi(p)\ge p\\ 0 & \text{otherwise.}\end{cases} \tag{20}$$

Twice more importance is given to deviations when $\xi(p)<p$ (inaccurate case) since then $|3a(p)-2|=2$. In the case where the fraction of true values falling

into the p-probability interval is larger than expected, i.e. the accurate case, this weight becomes 1. In other words, one penalizes the situation where the fraction of true values within the PI is smaller than expected. The ideal situation is when the experimental fractions match the theoretical ones, that is $G = 1$. The goodness statistics is completed by the so-called "accuracy plot" which is a plot of the experimental vs. expected fractions.

A cross-validation approach has been used to derive ccdfs of Cd, soil organic matter and clay content at, respectively, $N = 1553$, 1378 and 314 data locations of Fig. 2. Probability intervals have been computed for increasing probability p and the proportions of true values falling into these PIs were computed according to Eq. (16). Fig. 7 shows that for Cd the dots plot on the 45° line, which indicates that the theoretical fractions are correctly predicted by our uncertainty models (G close to 1). Results are not as good for the two other

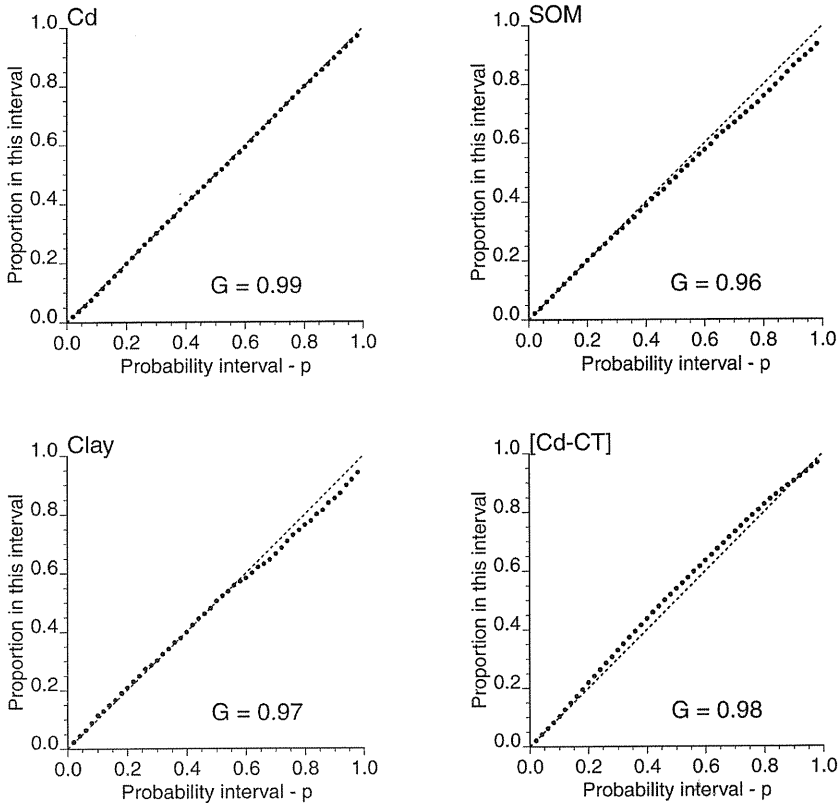

Fig. 7. Plots of the proportion of true values falling within probability intervals (accuracy plot) vs. the probability p. Ccdfs for Cd, organic matter and clay were derived using indicator kriging and cross-validation, while the ccdfs of the difference between Cd and the CT were obtained numerically using the Latin hypercube sampling procedure described in Fig. 8.

properties: the dots lying below the 45° line for $p > 0.5$ means that the proportion of true values falling into large p-probability intervals is smaller than expected, which reflects a less accurate modeling of the tails of the ccdfs. Deviations between experimental and theoretical fractions are however small since the goodness statistics are all very close to 1.

3.4. Latin hypercube sampling

3.4.1. Sanitation threshold

At each of the 2925 grid nodes, the ccdfs of soil organic matter and clay were discretised into 100 equiprobable classes which were randomly sampled each once and independently for both variables (the correlation coefficient between the 314 clay and soil organic matter values of the Belgian Soil Survey was 0.173, and their correlation should be even smaller on vegetable gardens due to the strong impact of gardening practice on soil organic matter). This yielded, for each grid node, 10 000 pairs of soil organic matter and clay values which were combined into Eq. (15) to generate a set of 10 000 CT values. Fig. 8 illustrates this procedure for one location. Fig. 9 shows a map of the mean of the local distributions (Eq. (13)) of the 10 000 CT values. The spatial variability of soil organic matter and clay, together with the analytical uncertainty, yielded CT values ranging between 1.60 and 2.51 mg kg^{-1}, with an average of 2.08 mg kg^{-1}.

3.4.2. Probability to exceed the CT

At the same 2925 grid nodes, the ccdf of the Cd content was also sampled by a Latin hypercube sampling (Fig. 8). The resulting set of 100 Cd values was compared with the set of 10 000 CT values yielding a ccdf based on 1 000 000 differences between Cd and CT. The underlying assumption here is the independence between CT and Cd values, the correlation of which could not be formally assessed since there is no location where all variables are known. In presence of a likely positive correlation between Cd and ST, an independent sampling of their probability distributions as performed in this paper would entail an overestimation of the actual risk of exceeding the CT (conservative scenario). Note that if the correlation can be estimated, procedures exist to sample jointly probability distributions so that the simulated values reproduce the experimental correlation (Heuvelink, 1998). The probability of exceeding the threshold was estimated by the proportion of these differences that are positive (Fig. 8) and these probabilities are mapped in Fig. 10. The probability of exceeding the CT varied between 0 and 1 with a mean of 0.62. As expected, high probabilities were found around the factories, but due to the spatial variability of all three variables involved, the spatial pattern of the probability to exceed the CT is much more complex than that of Cd (compare to Fig. 6).

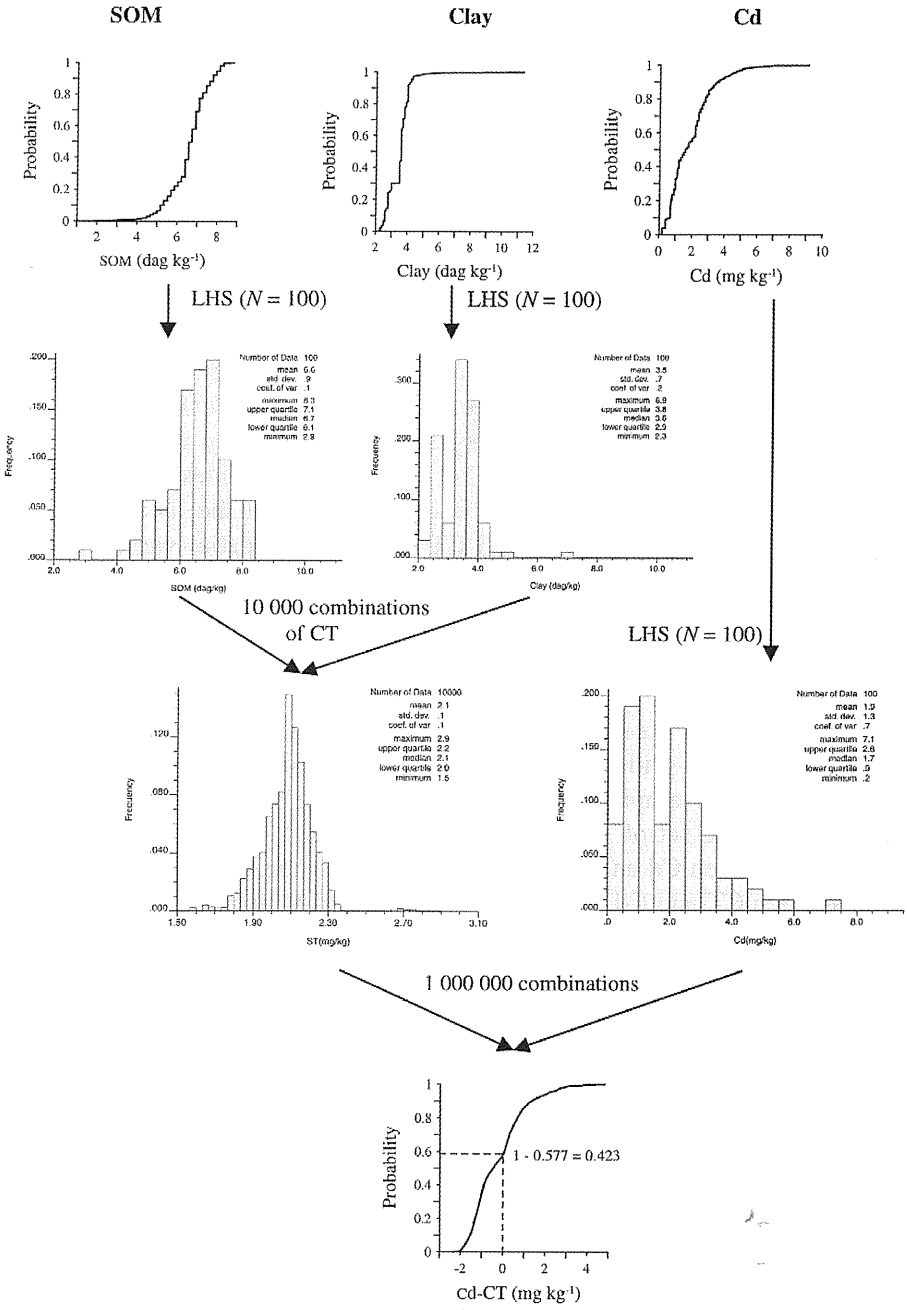

Fig. 8. Flowchart illustrating the calculation of the probability that the CT is exceeded at one location (215 900 m L72, 213 700 m L72) (LHS = Latin hypercube sampling).

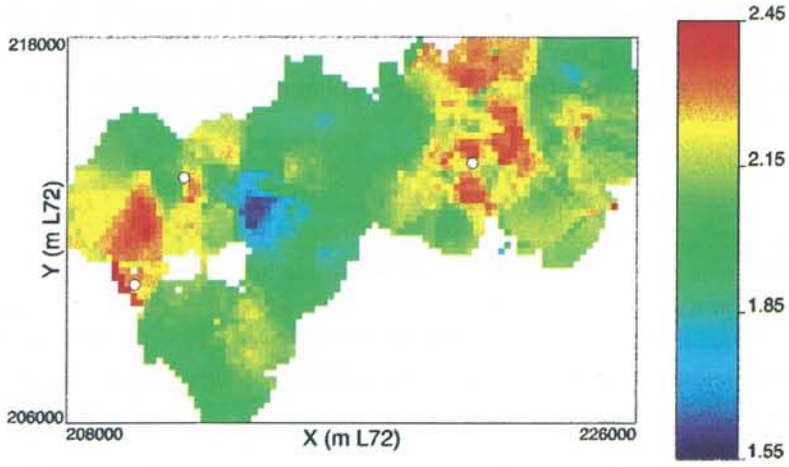

Fig. 9. Map of the mean of the local distribution of 10000 CT (mg kg^{-1}) values obtained through a Latin hypercube sampling of the ccdfs of soil organic matter and clay (circles locate factories).

The goodness of the ccdfs of the difference between Cd and CT was assessed using a cross-validation approach similar to the one described above. A difficulty was that such a cross-validation requires the knowledge of all three soil properties at the same locations, which is not the case here. Such a limitation was overcome by using only those 1364 locations where both Cd and organic matter content have been measured and by interpolating clay content at these locations using ordinary kriging. The later estimates have been considered as true values in the cross-validation approach. At each location, the three ccdfs have been sampled using the procedure described in Fig. 8 to yield a numeri-

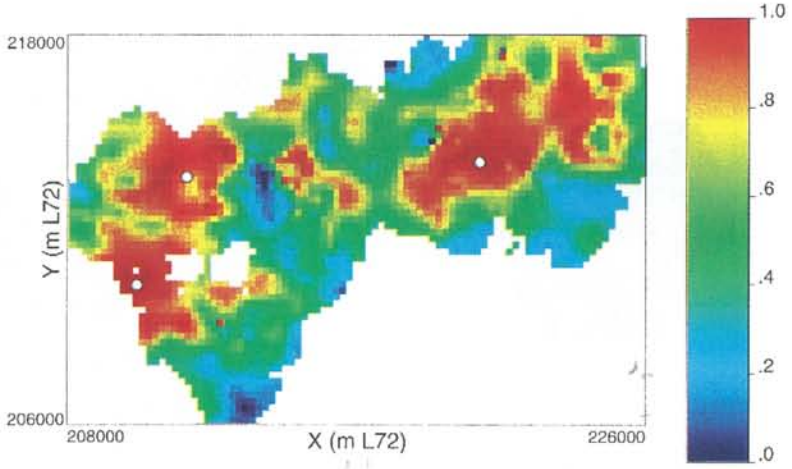

Fig. 10. Probability map to exceed the local CT.

cally cross-validated ccdf of the difference between Cd and CT. The accuracy plot of Fig. 7 (right bottom graph) shows that these models predict very well the proportions of true values that fall within probability intervals of increasing size ($G = 0.98$).

3.5. Advantages of the proposed procedure

The following are the advantages of our procedure.

(1) Any probability threshold p_c can be considered in decision-making. For example, if we would consider the 80% probability of exceeding CT as a critical probability level p_c, then $\mathbf{x}_0$ would be classified as contaminated if:

$$\text{Prob}\{Z(\mathbf{x}_0) > z_c|(n)\} \geq p_c. \tag{21}$$

In our case study, if $p_c = 0.80$ this would result in a classification of 27.3% of the interpolated area (covering 3192 ha) as unsafe to be used as vegetable gardens, i.e. where Cd exceeds the CT. To avoid the difficult selection of a probability threshold for criterion (21), an alternative consists of classifying as hazardous all locations where the CT is exceeded in expected value:

$$\mathbf{x}_0 \text{ is hazardous if } E[\text{Cd}(\mathbf{x}_0)] > E[\text{CT}(\mathbf{x}_0)] \text{ or if } E[\text{Cd}(\mathbf{x}_0) - \text{CT}(\mathbf{x}_0)] > 0. \tag{22}$$

The expected value of the difference between the Cd and CT values was numerically approximated by the arithmetic average of the 1 000 000 differences generated by the Latin hypercube sampling of the ccdfs. According to this criterion, the largest part of the interpolated area (72.7% or 8492 ha) was classified as hazardous (Fig. 11).

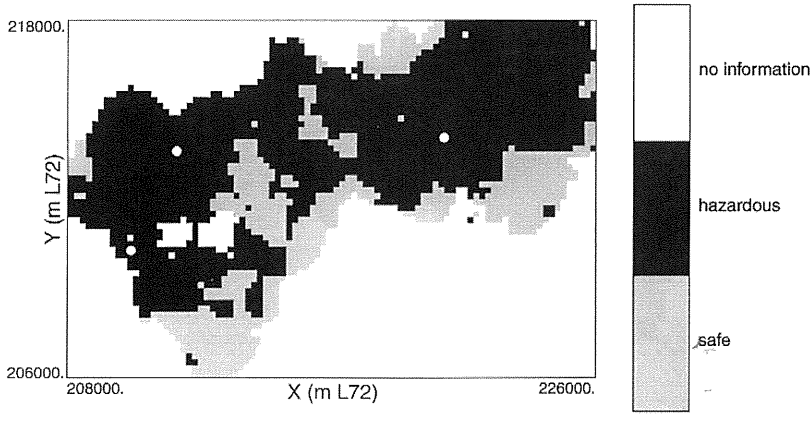

Fig. 11. Classification of locations as hazardous or safe on the basis that the Cd sanitation threshold is exceeded in expected value (Eq. (22)).

(2) Because the decision rule (22) involves expected values, there is actually a risk that the location $\mathbf{x}_0$ is wrongly declared hazardous (false positive) or safe (false negative). These can be computed directly from our results (Fig. 10) (Goovaerts, 1997; Myers, 1997).

(3) Using the concept of cost functions, an evaluation of the economic costs involved can be performed (Goovaerts et al., 1997).

An additional, but less often explored, advantage of our procedure is the ability to optimise the location of additional samples.

4. Location of additional samples

When we are very uncertain about the actual pollutant concentrations and the resulting risk of misclassification, it might be more advantageous to collect additional information before a final classification into safe or contaminated areas is being conducted (Van Groenigen, 1999). Consider that additional resources are available, allowing the collection of S additional samples and the measurement of Cd concentration, organic matter and clay contents. To increase the cost-effectiveness of the new sampling phase, it is important to account for the information already collected and processed.

4.1. Reduction of the uncertainty about Cd concentration

A first objective may be the reduction of the uncertainty about the Cd concentration, which is achieved by sampling the S locations with the largest ccdf variance for Cd:

$$\mathbf{x} \text{ is sampled if } s^2(\mathbf{x}) = \int_0^{+\infty} \{z - z_{\mathrm{E}}^*(\mathbf{x})\}^2 f(\mathbf{x}, z|(n))\mathrm{d}z \text{ is large} \tag{23}$$

where $f(\mathbf{x}; z|(n))$ is the local probability density function of $Z(\mathbf{x})$, and $z_{\mathrm{E}}^*(\mathbf{x})$ is the E-type estimate of $Z(\mathbf{x})$ (Eq. (13)). Fig. 12 (left column, top map) shows the map of the ccdf variance for Cd, and the 200 locations with the largest variance (left middle map). Because of heteroscedasticity (i.e. relation between local mean and local variance), the selected locations are all in the vicinity of the three factories. Moreover, they are strongly clustered, hence additional constraints must be imposed to avoid the collection of redundant information of these spatially autocorrelated variables. Now assume that $S = 50$ and that we accept 500 m as a maximal autocorrelation length (as a compromise between the different ranges of the indicator variograms, see Table 1). With this additional information we made a selection by starting from the location with the largest variance and remove all other location within 500 m. Next, the location with the second largest variance was selected and again all locations within 500 m were

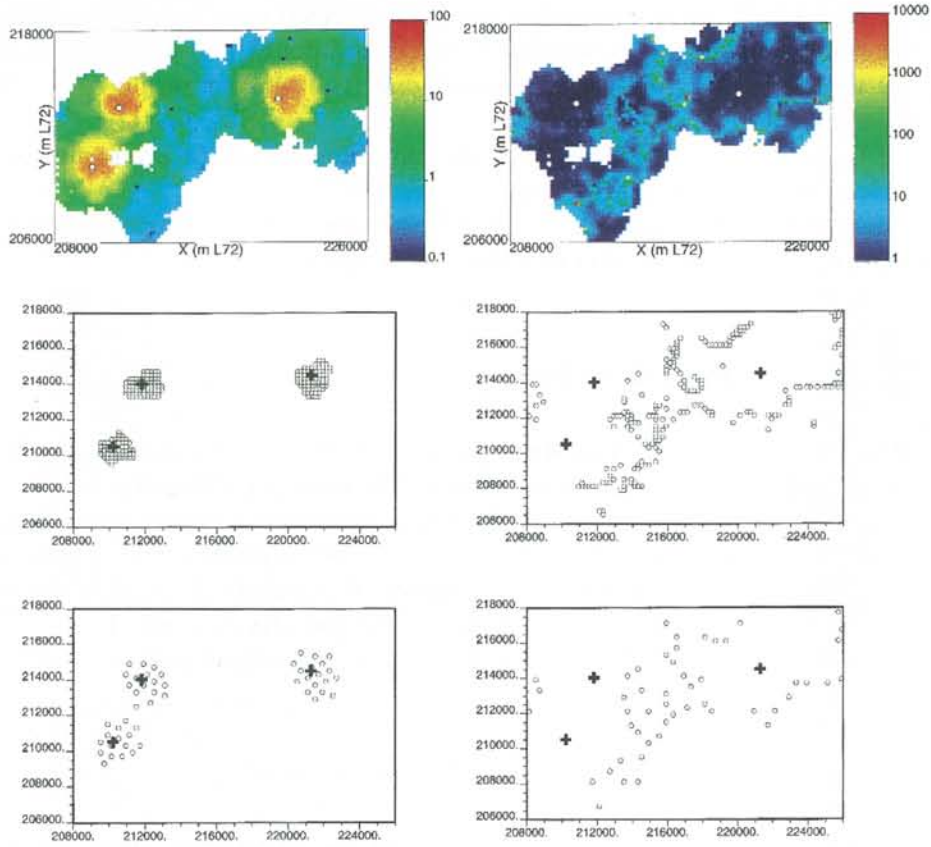

Fig. 12. Top left: map of ccdf variance for Cd ((mg kg^{-1})^2); top right: map of CV(**x**) (Eq. (24)); location of 200 locations candidate for additional sampling because of the large ccdf variance for Cd (left middle) or large CV(**x**) (right middle); bottom graphs show 50 locations with large ccdf variance for Cd (left) or large CV(**x**) (right) that are at least 500 m apart to increase the efficiency of the sampling design (circles in top maps and crosses in middle and bottom graphs locate factories).

removed. This procedure continued until 50 locations were obtained with the largest variance and that are at least 500 m apart, which increases the efficiency of the sampling. Fig. 12 (left bottom map) shows the result.

4.2. Reduction of the remediation error

In many situations, the primary concern is to assess the intensity and real extent of pollution concentrations exceeding a regulatory threshold. In other words, it is the remediation decision that matters, not the accuracy of the prediction itself (Rautman, 1997). The objective would then be to minimise the uncertainty about whether the critical threshold z_c is exceeded.

Chien (1998) proposed assessing the uncertainty at the unsampled location $\mathbf{x}$ by the product $\omega(\mathbf{x})s^2(\mathbf{x})$, where $s^2(\mathbf{x})$ is the conditional variance, a measure of local uncertainty as defined in Eq. (23) and $\omega(\mathbf{x})$ measures how close the probability of exceeding the critical threshold at $\mathbf{x}$ is to the probability threshold p_c specified by the decision-maker. Garcia and Froidevaux (1997) used as a measure of uncertainty the absolute difference between the probability of exceeding the critical threshold z_c and the closest of the two low and high risk probability thresholds 0.2 and 0.8; they considered that the uncertainty is negligible at locations where the probability of contamination is either high (> 0.8) or low (< 0.2). None of these methods take into account the fact that the predictions of the site-specific physical threshold z_c may be uncertain, as in our case study. Therefore, we propose to use as a sampling criterion the ratio of the standard deviation to the absolute value of the mean of the local cumulative distribution of the difference ($D(\mathbf{x})$) between the pollutant concentration ($Z(\mathbf{x})$) and the threshold ($Z_{CT}(\mathbf{x})$):

$$\mathbf{x} \text{ is sampled if } \mathrm{CV}(\mathbf{x}) = \frac{\sqrt{\mathrm{Var}[D(\mathbf{x})]}}{|E[D(\mathbf{x})]|} \text{ is large} \tag{24}$$

The expected value and the variance of $F_D(\mathbf{x};d|(n))$ were approximated by the arithmetical average and variance of the 1 000 000 differences between Cd and CT generated by the Latin hypercube sampling of the ccdfs. This type of coefficient of variation (CV) is large if the denominator is small, that is if the simulated pollutant concentrations and threshold values are close and so the uncertainty about the exceedence of that threshold becomes large. For the same average difference, the CV will be larger if the variance of the distribution of differences is large.

Fig. 12 (right column) shows the map of CV($\mathbf{x}$) (top), and the 200 locations with the largest values are displayed below it. Unlike the previous criterion, additional samples are no longer collected in the vicinity of factories, which is certainly contaminated, but the focus is on the borderline between the zones that could be classified as safe or hazardous (compare to Fig. 11). Indeed, it is in these areas that the risk for misclassification is the largest. Although the clustering is less pronounced than for the first criterion, the efficiency of the sampling can be increased by imposing a constraint of minimum distance between two samples; for example a minimum distance of 500 m leads to the selection of the 50 locations displayed at the bottom of Fig. 12 (right graph), using the same selection procedure as described before. This selection could be optimised further using simulated annealing (Van Groenigen and Stein, 1998); for example, the two constraints of maximisation of CV($\mathbf{x}$) and maximisation of the distance between samples could be included into a single objective function instead of imposing a constraint of minimum distance a posteriori (two-step approach).

5. Conclusions

The environmental database of our study consisted of three soil variables: Cd, soil organic matter and clay. These data had several characteristics complicating their combined spatial evaluation:

- Soil organic matter was collocated with Cd, but clay was not.
- Cd and soil organic matters were strongly spatially clustered, but only for Cd high-valued areas were preferentially sampled. A declustering procedure was used to obtain a Cd distribution that is representative of the study area.
- The distributions of these variables were moderately (soil organic matter and clay) to strongly (Cd) positively skewed.
- Cd displayed a strong spatial trend due to the preferential winds and the distance to its sources (three factories).
- The analytical repeatability varied between 2.2% (for soil organic matter) to 7.8% (for Cd).

The aim of this paper was to present a non-parametric methodology to incorporate two common sources of uncertainty, spatial interpolation and analytical error, into the prediction of the probability of exceeding a location-specific threshold. The combination of (non-stationary in the case of Cd) indicator kriging and Latin hypercube sampling yielded local ccdfs of the difference between the heavy metal concentration and the CT. Knowledge of such ccdfs allowed the mapping of the probability of exceeding that threshold, i.e. in our case study this amounted to 27.3% of the interpolated area (3190 ha) where the probability to exceed the CT is 80% or higher. Cross-validation results indicated that most ccdf models are accurate in that the fraction of true values falling into a p-probability interval is usually larger than expected.

The design of a sampling scheme that minimises the averaged kriging variance over the study area typically leads to take additional samples in sparsely sampled areas. Although it is important to account for first-phase sampling density in the elaboration of the second-phase design, data values must also be accounted for, in particular in the presence of heteroscedasticity. In this case, the uncertainty may be larger in an area that is densely sampled but which displays higher variabilities than in a sparsely sampled area that is homogeneous. Whereas minimisation of uncertainty about Cd concentration entails the sampling of high-valued areas around factories, the proposed approach (i.e. minimisation of the "coefficient of variation" of the ccdf of differences between Cd concentration and the CT) leads to the sampling of borderlines between areas classified as hazardous or safe. Because the ccdfs are data-dependent, one cannot investigate a priori the impact of the sampling strategy on the local uncertainty, as is possible when the kriging variance is used as a measure of uncertainty (Burgess et al., 1981). Thus, additional constraints, such as mini-

mum distance between samples, need to be imposed to avoid clustering of samples and the consequent loss of efficiency of the sampling design.

Acknowledgements

J. Hendrickx (community of Balen), G. Ide (LISEC), W. Mennen (community of Lommel), A. Stein (Landbouwuniversiteit Wageningen), J. Vangronsveld (Limburgs Universitair Centrum), and D. Wildemeersch (Vlaamse Gemeenschap) are kindly thanked for providing the Cd and soil organic matter data. This research was conducted during a leave of the first author at The University of Michigan for which he gratefully acknowledges the financial support of the Flemish Fund for Scientific Research (FWO). The authors also wish to express their gratitude to the reviewers for their constructive remarks.

References

Burgess, T.M., Webster, R., McBratney, A.B., 1981. Optimal interpolation and isarithmic mapping of soil properties: IV. Sampling strategy. Journal of Soil Science 31, 315–331.

Burrough, P.A., McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford Univ. Press, Oxford, 333 pp.

Chaney, R.L., 1990. Public health and sludge utilization, part II. BioCycle 31, 68–73.

Chien, Y.-J., 1998. A Geostatistical Approach to Sampling Design for Contaminated Site Characterization. Master Thesis, Stanford University, Stanford, CA.

Deutsch, C.V., 1997. Direct assessment of local accuracy and precision. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics Wollongong '96. Kluwer Academic Publishing, Dordrecht, pp. 115–125.

Deutsch, C.V., Journel, A.G., 1998. GSLIB Geostatistical Software Library and User's Guide. Oxford Univ. Press, New York.

Englund, E.J., Heravi, N., 1994. Phased sampling for soil remediation. Environmental and Ecological Statistics 1, 247–263.

Garcia, M., Froidevaux, R., 1997. Application of geostatistics to 3-D modelling of contaminated sites: a case-study. In: Soares, A. (Ed.), geoENV I—Geostatistics for Environmental Applications. Kluwer Academic Publishing, Dordrecht, pp. 309–325.

Goovaerts, P. et al., 1994. Comparative performance of indicator algorithms for modeling conditional probability distribution functions. Mathematical Geology 26, 389–411.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York, 483 pp.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45.

Goovaerts, P., Journel, A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. European Journal of Soil Science 46, 397–414.

Goovaerts, P., Webster, R., Dubois, J.-P., 1997. Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environmental and Ecological Statistics 4, 31–48.

Heuvelink, G.B.M., 1998. Error Propagation in Environmental Modeling with GIS. Taylor, Francis, London, 127 pp.

Journel, A.G., 1983. Non-parametric estimation of spatial distributions. Mathematical Geology 15, 445–468.

Journel, A.G., 1986. Constrained interpolation and qualitative information. Mathematical Geology 18, 269–286.

Journel, A.G., Huijbregts, A.G., 1978. Mining Geostatistics. Academic Press, New York, 600 pp.

Juang, K.-W., Lee, D.-Y., 1998. A comparison of three kriging methods using auxiliary variables in heavy-metal contaminated soils. Journal of Environmental Quality 27, 355–363.

Luxmoore, R.J., King, A.W., Tharp, L., 1991. Approaches to scaling up physiologically based soil-plant models in space and time. Tree Physiology 9, 281–292.

McBratney, A.B., Webster, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. Journal of Soil Science 37, 617–639.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Myers, J.C., 1997. Geostatistical Error Management. Van Nostrand-Reinhold, New York, 571 pp.

OFEFP (Office fédéral de l'environnement, des forêts et du paysage), 1993. Réseau national d'observation des sols, période d'observation 1985–1991. Cahier de l'environnement no. 200, Berne, Swiss.

Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics (accepted for publication).

Rautman, C.A., 1997. Geostatistics and cost-effective environmental remediation. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics Wollongong '96. Kluwer Academic Publishing, Dordrecht, pp. 941–950.

Tiktak, A., Leijnse, A., Vissenberg, H., 1999. Uncertainty in a regional-scale assessment of cadmium accumulation in the Netherlands. Journal of Environmental Quality 28, 461–470.

Van Groenigen, J.W., 1999. Constrained optimisation of spatial sampling—a geostatistical approach. PhD Thesis, Wageningen Agricultural University and ITC Enschede, 148 pp.

Van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27, 1078–1086.

Van Meirvenne, M., 1991. Characterization of soil spatial variation using geostatistics. PhD thesis, University of Gent.

Vlaamse Gemeenschap, 1996. Besluit van de Vlaamse regering houdende vaststelling van het Vlaams reglement betreffende de bodemsanering. Belgisch Staatsblad dd. 27.03.1996, pp. 7018–7058.

Webster, R., Burgess, T.M., 1984. Sampling and bulking strategies for estimating soil properties of small regions. Journal of Soil Science 35, 127–140.