Geoderma 158 (2010) 242–251

Contents lists available at ScienceDirect

Geoderma

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

Delineating the hazard zone of multiple soil pollutants by multivariate indicator kriging and conditioned Latin hypercube sampling

Hone-Jay Chu [a], Yu-Pin Lin [a,*], Cheng-Shin Jang [b], Tsun-Kuo Chang [a]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, 1, Sec. 4, Roosevelt Rd., Da-an District, Taipei City 106, Taiwan, R.O.C.
[b] Department of Leisure and Recreation Management, Kainan University, No.1 Kainan Road, Luzhu, Taoyuan 33857, Taiwan, R.O.C.

ARTICLE INFO

Article history:
Received 29 October 2009
Received in revised form 27 April 2010
Accepted 3 May 2010
Available online 9 June 2010

Keywords:
Conditioned Latin hypercube sampling
Multiple-variable indicator kriging
Heavy metal
Soil contaminant

ABSTRACT

Soil pollution delineation is a difficult task because there may be multiple pollution sources and complex transport schemes. In this study, four heavy-metal concentrations, namely Cr, Cu, Ni, and Zn, were collected at 1082 sampling sites in central Taiwan. Conditioned Latin Hypercube Sampling (cLHS) is used to determine the most cost-effective sampling strategies for long-term monitoring of multiple heavy metals. Then, the study estimates the probability of multiple hazardous heavy metals using multiple-variable indicator kriging (MVIK) based on sufficient cLHS samples.

The results suggest that heavy-metal sampling patterns, including the size and configuration, affect the spatial distribution of probable hazards, especially when the sample size is small. Unlikely random sampling, the cLHS method replicates the variability and distribution of variables. In this study, an area is defined as a hazard zone if the amount of any heavy metal exceeds the corresponding regulatory threshold. The heavy-metal delineations (e.g., where the MVIK hazard probability exceeds 0.85) cover over 20% of the study area and correlate highly with the locations of industrial plants and irrigation systems in the area. Hence, MVIK coupled with cLHS provides a way to assess the presence of multiple hazards efficiently and effectively in future monitoring and environmental management projects.

1. Introduction

Soil contamination is caused by the presence of man-made chemicals or other alterations in natural soil environment. The common sources of chemicals are petroleum hydrocarbons, pesticides, solvents, and heavy metals. The occurrence of this phenomenon is highly correlated with the degree of industrialization and intensity of chemical usage (Lin, 2002; Franco et al., 2006; Zhao et al., 2007). The concern over soil contamination stems primarily from health risks it poses through direct contact and from secondary contamination of water supplies.

However, the delineation of soil pollution is usually a challenging task, especially when there are multiple pollution sources and the composition of the polluting effluents is complex (Goovaerts, 2001). Soil pollution data for heavy metals occasionally exhibits small-scale discontinuities or variations that increase the difficulty of delineating the characteristics of soil pollutants (Lin, 2008). The data may be sparely distributed across the sample sites, so it cannot provide all the information needed for risk assessment and effective environmental management. Moreover, the collected data may be inconsistent, possibly because of complicated temporal and spatial variations in the measurable characteristics of the investigated pollution sources.

In the delineation of contamination areas, risk assessment at unsampled locations is significantly important (Goovaerts, 2001; Van Meirvenne and Goovaerts, 2001; Lin, 2002; Schnabel, 2004; Amini, 2005; Hassan and Atkins, 2007; Lin, 2008). Moreover, in an urbanized, industrialized or agricultural area, the soil might be polluted by several heavy metals (Lin et al., 2002; Krishna and Govil, 2004; Norra et al., 2006; Lin et al., 2010). Mapping multiple-contaminant polluted areas requires the use of multivariate analysis with the spatial estimation (Smith et al., 1993; Diodato and Ceccarelli, 2004; Juang et al., 2004; Zhang et al., 2008). Multiple-variable indicator kriging (MVIK) is a multivariate approach that provides more conservative information and scenarios than single-variable indicator kriging (Jang et al., 2008; Lee et al., 2008; Lin et al., 2010). Furthermore, hazardous probability maps for multiple pollutants are needed for soil investigations. Many studies have applied MVIK successfully in investigations of soil quality in farmed fields (Smith et al., 1993; Halvorson et al., 1996; Oyedele et al., 1996; Diodato and Ceccarelli, 2004). However, few studies have used MVIK to obtain representative samples to identify the hazard zone of contaminants.

According to the EPA's guidelines for collecting environmental data, soil samples can be gathered by judgmental sampling, random sampling, stratified sampling, or systematic sampling. Because of the complicated distributions of soil contaminations, random sampling

* Corresponding author. Tel.: +886 2 33663467; fax: +886 2 33663464.
E-mail addresses: honejaychu@gmail.com (H.-J. Chu), yplin@ntu.edu.tw (Y.-P. Lin), csjang@mail.knu.edu.tw (C.-S. Jang), tknchang@ntu.edu.tw (T.-K. Chang).

0016-7061/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2010.05.003

may not be efficient for delineating the hazard zone of contaminants. Hence, other sampling designs, such as stratified sampling and systematic sampling, may be preferred for analysis, but the sampling procedure is complicated because of the nature of the contaminant data. Moreover, long-term monitoring of a polluted area is necessary to ensure that remediation objectives are achieved to reduce the risk to people's health and the environment. Several hundred samples must be collected and analyzed each year, but that incurs numerous costs. Reducing the sample size could lower the monitoring costs. Thus, for reasons of cost-efficiency, it is better to select sites that will make the greatest contribution to the characterization of contaminant over the entire domain. Latin hypercube sampling (LHS), a stratified random procedure, is an efficient means of sampling from multivariate distributions (McKay et al., 1979; Minasny and McBratney, 2006). A number of studies posit that LHS can explore a model's parameter space more exhaustively than simple random sampling (SRS) (McKay et al., 1979; Xu et al., 2005; Hassan and Atkins, 2007; Post et al., 2008). LHS yields a more representative distribution of model outputs given the same number of input simulated realizations. The efficiency of LHS is better than that of SRS for univariate distributions (McKay et al., 1979). Post et al. (2008) discussed the origins of a model's uncertainty in terms of the model's inputs. They used the Monte Carlo method with LHS to investigate temporal propagation and river basin scale propagation of uncertainty in long-term soil organic carbon (SOC) dynamics. Xu et al. (2005) combined stochastic simulation and LHS in a forest landscape simulation model. Their results show that LHS can capture more variability in the sample space than SRS, especially when the number of simulations is small.

Developing efficient procedures for collecting information-effective samples is essential if we gain a more accurate understanding of the spatial distribution of pollutants. Carre et al. (2007) proposed using LHS as a sampling design for digital soil mapping. They used the principle of hypercube sampling to assess the quality of existing soil data and identified the locations that needed to be sampled. The proposed algorithm checks the occupancy of legacy sampling units in the hypercube of the quantiles of the environmental data. However, in practice, many surveys consider more than one variable in soil sampling. A multivariate sampling strategy is essential. Minasny and McBratney (2006) employed conditioned Latin hypercube sampling (cLHS) with prior information to represent ancillary data for the Hunter Valley in New South Wales, Australia. They also used the cLHS approach with a search algorithm based on annealing schedules derived from the multivariate distributions. The results show that cLHS is more effective than SRS and stratified spatial sampling for replicating the distribution of variables. The cLHS method provides full coverage of each variable by maximally stratifying the marginal distribution and ensuring a good spread of sampling points (Minasny and McBratney, 2006; Lin et al., accepted for publication).

The primary objective of the present work is to investigate proposals for sampling and delineating designs for multiple-contaminant management. In this study, cLHS and MVIK are combined to determine the hazard probability of multiple pollutants. First, cLHS ensures the correlation of the sampled variables used to replicate the original data. Then, MVIK generates the probability map of multiple hazards based on the cLHS samples. A study case from the field survey is provided through comparison of modeling with SRS and conditional simulations that estimate the contaminant maps and hazard delineation.

2. Materials and methodology

2.1. Study area

The research area is in the northern part of Chanhua County, Taiwan. Changhua City lies to the east and Lugang Township lies to the west. The area, which is covered by irrigation systems, is regarded as an important agricultural region (Fig. 1). The study area includes suspected pollution sources and many factories such as metal work, dyeing and finishing, surface treatment, livestock, textile, electroplating, and tannery businesses (Lin et al., 2010). In this study area, the data of 1082 topsoil (0–15 cm) samples containing concentrations of Cr, Cu, Ni, and Zn were obtained by the soil heavy-metal investigation project carried by Taiwan's Environmental Protection Administration (EPA), between February and August 2002. The sampling sites are also shown in Fig. 1. Approximately 1 kg of soil was collected for each sample using a stainless steel spade and a plastic scoop and then stored in a plastic food bag. After air drying at room temperature, 3 g of each soil sample was disaggregated, sieved to 0.85 mm and ground to a fine 0.15 mm powder. Each 3 g milled sample was then digested for 2 h at room temperature with 7 mL HNO_3 and 21 mL HCl (aqua regia, 1:3) to slowly oxidize organic matter in the soil. Next, the digest, 100 mL aqua regia, was filtered. Additionally, the levels of Cr, Cu, Ni, and Zn in the samples were determined by Inductively Coupled Plasma-Optical Emission Spectrometers (ICP-OES) (Lin et al., 2010).

2.2. Multiple-variable indicator kriging (MVIK)

MVIK is basically the same as ordinary kriging, but the binary data is generated by a thresholding procedure (Smith et al., 1993; Jang et al., 2008). Indicator kriging provides a non-parametric distribution estimated directly using fixed thresholds by considering indicator transforms of the conditioning data in the form of cumulative distribution functions with step functions. The method estimates the probability that the concentration of a pollutant exceeds a specific threshold value at a given location (Deutsch and Journel, 1992; Lin, 2002). The data ($Z(x)$) is transformed into an indicator as follows:

$$i(x, z_c) = \begin{cases} 1, & \text{if } Z(x) \le z_c \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

If the concentration of heavy metal ($Z(x)$) exceeds z_c, the indicator will be 0; otherwise it will be 1 (Goovaerts et al., 1997).

For MVIK, the integration of K heavy-metal indicators is achieved by the intersection operator, and is defined as follows:

$$I(x, z_c) = \min\left(i_k\left(x, z_c^k\right)\right), k = 1, \ldots, K. \tag{2}$$

The expected value of $I(x;\, z_c|(n))$, conditional on n surrounding data, can be expressed as

$$E[I(x; z_c\,|\,(n))] = \Pr ob[Z(x) \le z_c\,|\,(n)]. \tag{3}$$

The hazard probability that exceeds z_c can be expressed as

$$\Pr ob[Z(x) > z_c\,|\,(n)] = 1 - \Pr ob[Z(x) \le z_c\,|\,(n)]. \tag{4}$$

Eqs. (2) and (4) imply that a hazard exists if any one of multiple heavy metals exceeds the corresponding threshold.

This ordinary indicator kriging estimator is defined as follows:

$$\Pr ob[z(x_0) \le z_c\,|\,(n)] = \sum_{\alpha=1}^{n} \lambda_\alpha I(x_\alpha; z_c), \tag{5}$$

where $I(x_\alpha;\, z_c)$ represents the indicator values at x_α; and $\alpha = 1, \cdots, n$; λ_α is the kriging weight of $I(x_\alpha;\, z_c)$ determined by solving the following kriging system:

$$\sum_{\beta=1}^{n} \lambda_\beta \gamma_i\left(x_\alpha - x_\beta; z_c\right) + \mu = \gamma_i(x_\alpha - x_0; z_c), \tag{6}$$

$$\sum_{\beta=1}^{n} \lambda_\beta = 1, \tag{7}$$

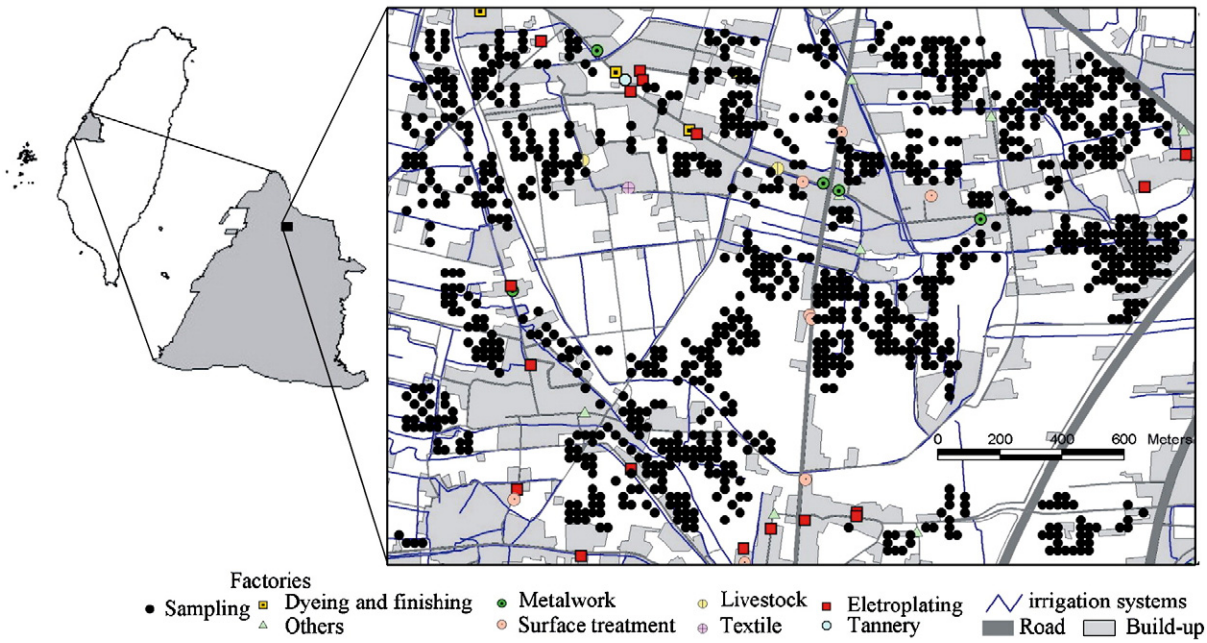

Fig. 1. The study area and sampling sites.

where μ is the Lagrange multiplier; $\gamma_i(x_\alpha - x_\beta;\ z_c)$ is the indicator variogram between the indicator variables at the αth and βth sampling points; $\gamma_i(x_\alpha - x_0;\ z_c)$ is the variogram between the indicator variables, i.e., the αth sampling point and x_0; and $\alpha = 1, \cdots, n$.

2.3. Sequential indicator simulation (SIS)

SIS is used for mapping, assessing and validating the hazard delineation based on 1082 samples. In the algorithm, modeling of the N-point conditional cumulative distribution function (ccdf) is a sequence of N univariate ccdfs at each grid cell along a random path (Goovaerts, 1997; Lin et al., 2009). The steps of the sequential indicator simulation algorithm are as follows (Deutsch and Journel, 1992; Goovaerts, 1997; Lin et al., 2009):

1. Establish a random path that visits each location of the domain once, where all nodes $\{x_i, i = 1, \cdots, N\}$ discretize the domain of interest. A random visiting sequence ensures that a spatial continuity artifact is not introduced to the simulation by a specific path visiting N nodes.
2. At the first visited node (x_1):
   A. Using either a parametric or non-parametric approach, model the local ccdf of $Z(x_1)$ conditional on n original data $\{Z(x_\alpha), \alpha = 1, \cdots, n\}$: $F_Z(x_1;\ z_1|(n).) = \text{prob}\{Z(x_1) \le z_1|(n)\}$.
   B. Via the Monte Carlo approach, generate, a simulated value $z^{(l)}(x_1)$ from the derived ccdf $F_Z(x_1;\ z_1|(n))$, and add it to the conditioning data set.
3. At the ith node x_i on the random path:
   A. Model the local ccdf of $Z(x_i)$ conditional on n original data and the closest $i-1$ previously simulated values $\{z^{(l)}(x_j), j = 1, \cdots, i-1\}$:

$$F_Z(x_i; z_i|(n + i - 1)) = \{Z(x_i) \le z_i|(n + i - 1)\}. \tag{8}$$

   B. Generate a simulated value $(z^{(l)}(x_i))$ from this ccdf, and add it to the conditioning data set, which is now of dimension $n + i$.
4. Repeat step 3 until all N nodes along the random path have been visited.

In SIS, the indicator kriging estimator is used to model the prior ccdf at each unsampled location (Juang et al., 2004). Since modeling the prior ccdf at each unsampled location should use values previously simulated at other sampled locations, the simulated values for all unsampled locations are referred to as a joint realization (Goovaerts, 1997; Juang et al., 2004). In this study, the cutoff values for each heavy metal are the 25th, 50th, and 75th percentiles of the prior ccdf.

2.4. Conditioned Latin hypercube sampling (cLHS)

In this paper, cLHS with a search algorithm based on heuristic rules is combined with an annealing schedule (Metropolis et al., 1953; Minasny and McBratney, 2006). The cLHS procedure represents the following optimization problem: given N sites with ancillary data (Z), select n sample sites $(n \ll N)$ such that the sampled sites form a Latin hypercube. For k continuous variables, each component of Z is divided into n (sample size) equally probable strata based on their distributions, and z is a sub-sample of Z. The steps of the cLHS algorithm (Minasny and McBratney, 2006; Lin et al., accepted for publication) are as follows:

1. Divide the quantile distribution of Z into n strata, and calculate the quantile distribution for each variable, $q_j^i, \ldots, q_j^{n+1}$. Calculate the correlation matrix of $Z(C)$.
2. Select n random samples from N, and calculate the correlation matrix of $z(T)$.
3. Calculate the objective function. The overall objective function combines different components of the objective function. For general applications, the weightings assigned to all the components of the objective function are equal.
   A. Since heavy-metal concentrations are continuous variables, an objective function is

$$O_1 = \sum_{i=1}^{n} \sum_{j=1}^{k} \left| \eta\left(q_j^i \le z_j \le q_j^{i+1}\right) - 1 \right|, \tag{9}$$

   where $\eta(q_j^i \le z_j \le q_j^{i+1})$ is the number of z_j that falls between quantiles q_j^i and q_j^{i+1}.

B. To ensure that the correlation of the sampled variables will replicate the original data, another objective function is defined:

$$O_2 = \sum_{i=1}^{k} \sum_{j=1}^{k} \left| c_{ij} - t_{ij} \right|, \tag{10}$$

where c denotes an element of C, the correlation matrix of Z; and t is the equivalent element of T, the correlation matrix of z.

4. Implement an annealing schedule: $M = \exp[-\Delta O/\mathrm{T}]$, where ΔO is the change in the objective function, and T is a cooling temperature (between 0 and 1), which is reduced by a factor in each iteration.
5. Generate a uniform random number between 0 and 1. If rand $< M$, accept the new values; otherwise, discard the changes.
6. Try to apply changes: Generate a uniform random number rand. If rand $< P$, pick a sample at random from z and swap it with a random site from the unsampled sites r. Otherwise, remove the sample(s) from z that has (have) the largest $\eta(q_j^i \le z_j \le q_j^{i+1})$ and replace it (them) with a random site(s) from unsampled sites r. End when the value of P is between 0 and 1, indicating that the probability of the search is a random search; otherwise, systematically replace the samples that have the worst fit with the strata.
7. Repeat steps 3–7 until the objective function value falls beyond a given stop criterion or a specified number of iterations are completed.

3. Results and discussion

3.1. Data statistics and spatial simulations

Fig. 2 illustrates the histograms of four heavy-metal pollutants and summarizes the descriptive statistics (such as mean and standard deviation) for the 1082 samples. The sample histograms of four pollutants are strongly and positively skewed. In this study, the probability of a contaminated area was mapped using MVIK with the regulatory thresholds. In Taiwan, the regulatory thresholds (maximum allowable concentrations) for the investigated heavy metals are as follows: Cr: 250 mg/kg, Cu: 200 mg/kg, Ni: 200 mg/kg, and Zn: 600 mg/kg. For the 1082 samples, the regulatory thresholds for Cr, Cu, Ni, and Zn are the 76th, 68th, 48th, and 71th percentiles, respectively. The statistical analysis shows that Ni contamination exceeds the regulatory threshold substantially and it is the most serious pollutant in the study area.

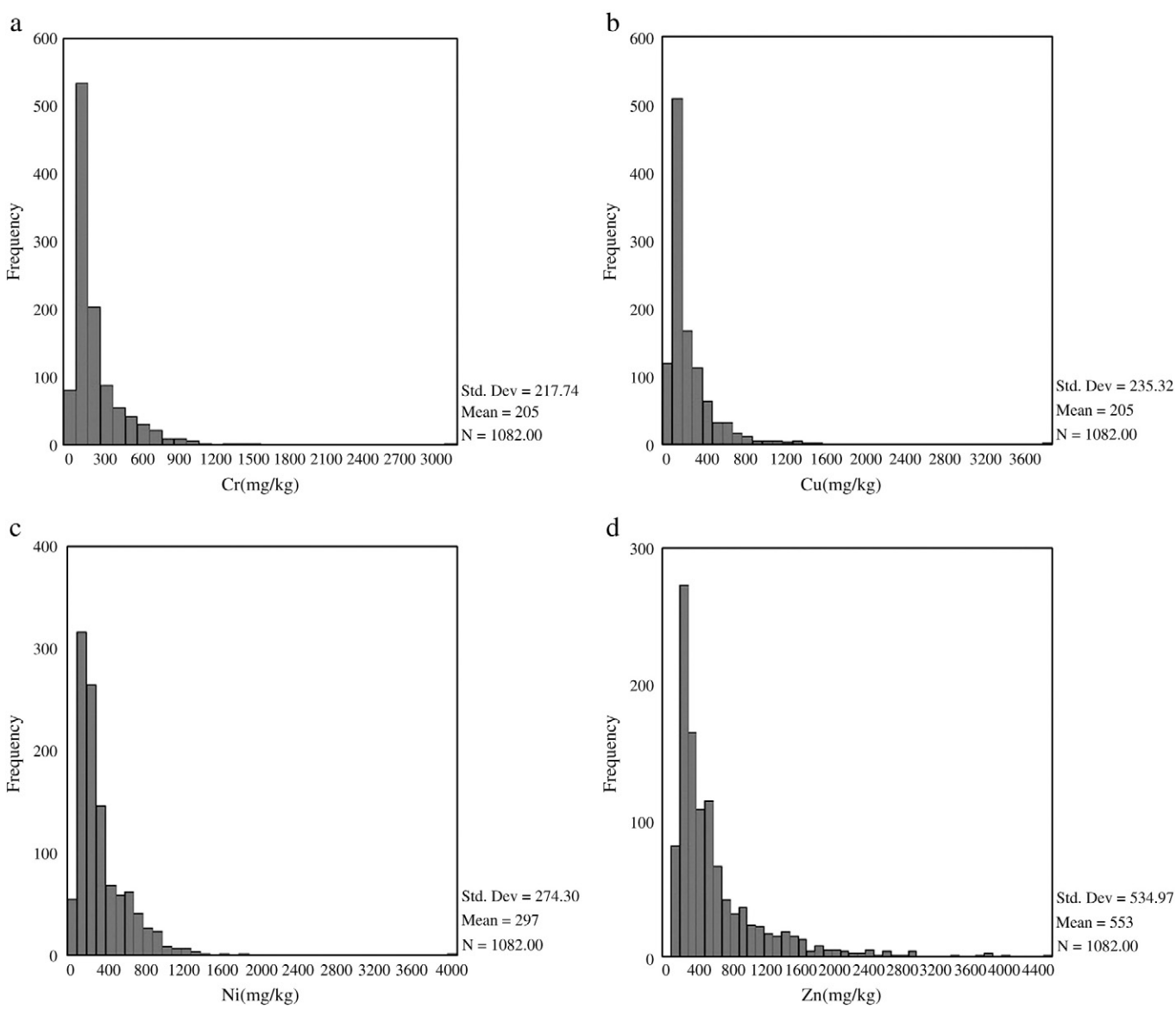

Fig. 2. Histograms for four pollutants (a) Cr, (b) Cu, (c) Ni, and (d) Zn in the study area.

Table 1
Indicator variogram models for the 25th, 50th, and 75th percentiles of heavy metals in 1082 samples.

Heavy metal		Model	Parameters			RSS	r^2
			C0	C0 + C	R		
Cr	25%	Exp.	0.020	0.184	216	1.730E-03	0.722
	50%	Exp.	0.026	0.247	171	1.202E-03	0.807
	75%	Exp.	0.025	0.190	120	2.075E-04	0.852
Cu	25%	Exp.	0.017	0.184	240	2.008E-03	0.737
	50%	Exp.	0.025	0.247	186	7.016E-04	0.899
	75%	Exp.	0.024	0.190	108	5.293E-04	0.663
Ni	25%	Exp.	0.015	0.179	222	2.614E-03	0.634
	50%	Exp.	0.022	0.237	228	3.608E-03	0.671
	75%	Exp.	0.018	0.183	159	5.723E-04	0.805
Zn	25%	Exp.	0.024	0.190	222	1.464E-03	0.768
	50%	Exp.	0.028	0.250	171	3.795E-04	0.936
	75%	Exp.	0.021	0.189	144	8.077E-03	0.710

Exp.: Exponential model; C0: Nugget; C0 + C: Sill; R: Range; RSS: Residual Sum of Squares; r^2: Coefficient of determination.

The Sequential Indicator Simulations were based on the indicator variogram models for the 25th, 50th, and 75th percentiles of the sample distribution of 1082 samples for Cr, Cu, Ni, and Zn in the study area (Table 1). The variograms of the four heavy-metal concentrations provide information about the spatial variability of soil heavy metals using GS+ (2004). Fig. 3 shows the average concentration of each heavy metal in 1000 SIS realizations based on the original data. Generally, SIS is preferred over interpolation algorithms for applications where the distribution of spatial variation is skewed. The reason is that SIS does not make any assumption about the shape of the conditional distribution (Deutsch and Journel, 1992; Goovaerts, 1997; Zhao et al., 2008). In the model, the dimensions are 72 (rows) by 104 (columns) and the size of each grid is 25 m by 25 m. The results demonstrate that the hotspots of Cr and Cu are similar. The spatial patterns also reveal high concentrations of Cr near industrial plants and irrigation systems in the study area. The Cu hotspots are located in the central and eastern parts of the study area in the vicinity of industrial plants and irrigation systems. The hotspots of Ni are distributed throughout the study area, except for the south-western

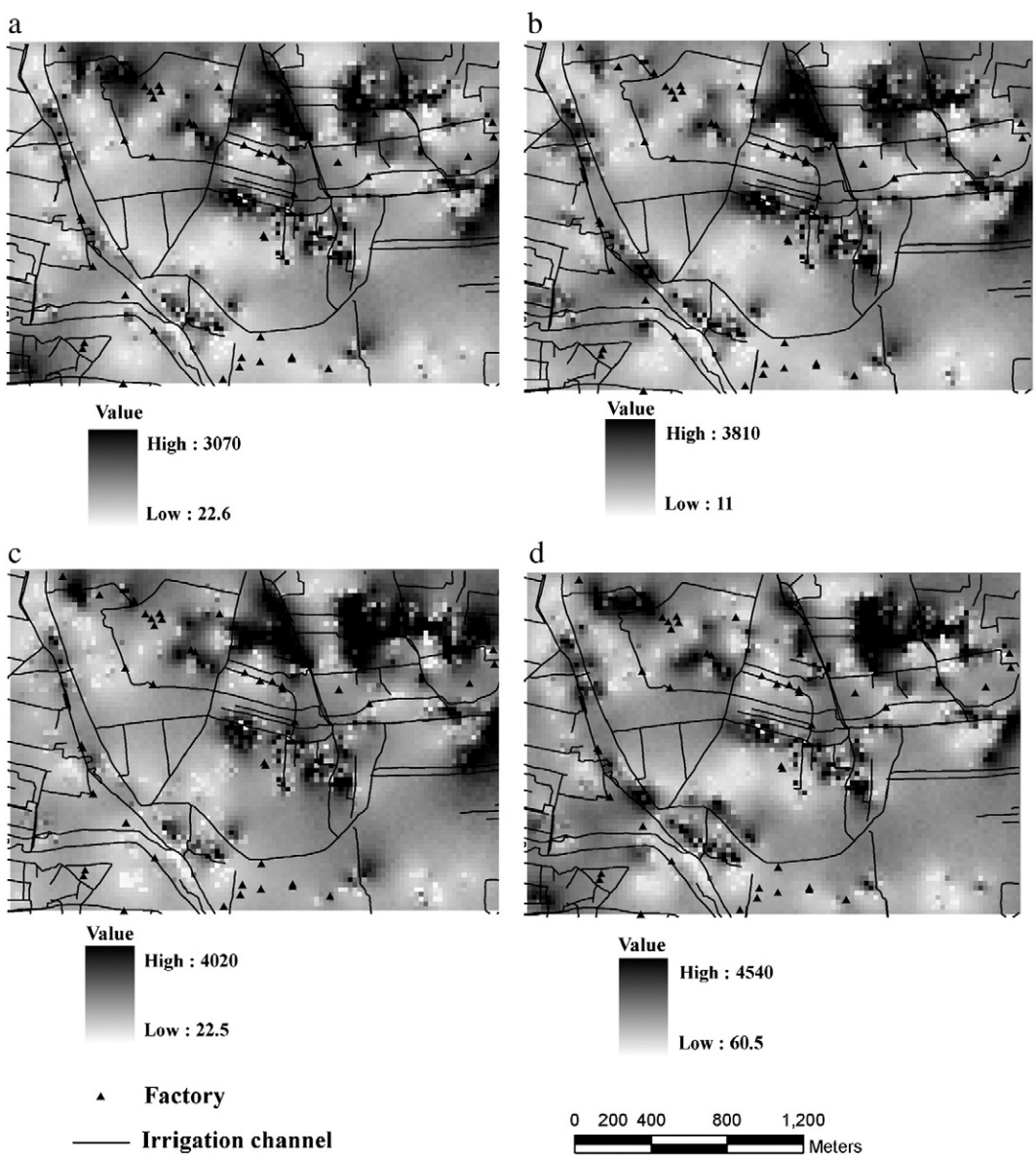

Fig. 3. The average concentration maps of (a) Cr, (b) Cu, (c) Ni, and (d) Zn in 1000 SIS realizations based on 1082 samples (unit: mg/kg).

part; and the areas with high concentrations of Zn are close to industrial plants and irrigation systems in the north-western part. Fig. 4 shows the probability maps for sites where Cr, Cu, Ni, and Zn exceed the regulatory thresholds based on the 1000 SIS realizations. Overall, the maps show that the highest variability is close to industrial plants and irrigation systems. The results, which match those of previous studies, show that the distributions of heavy-metal usage and pollution sources correlate with industrial plants and irrigation channels (Lin et al., 2010). As mentioned earlier, there are various factories in the study area, including metal work, surface treatment, livestock, textile, electroplating, dyeing and tannery plants; and they are suspected of discharging contaminated waste-water into the irrigation systems (Lin et al., 2002; Lin et al., 2010). Furthermore, soils are affected by a number of pollutants resulting from different human activities (Zhao et al., 2007). The occurrences of soil contamination are correlated with the degree of industrialization and the intensity of chemical usage, especially heavy metals (Lin et al., 2001; Lin, 2002; Krishna and Govil, 2004; Franco et al., 2006; Kasassi et al., 2008).

3.2. Hazard probabilities derived by MVIK based on 1082 samples

The indicator geostatistics allow a straightforward assessment of probabilities in excess of the critical values (Goovaerts, 1998). This work estimates the occurrence probabilities of the four hazardous metals based on MVIK. Fig. 5 (a) shows the MVIK probability map of the hazardous zone based on 1082 samples. MVIK could be used to assess the risk of exceeding the regulatory thresholds at unsampled locations, and to simulate the spatial distribution of hazardous zones for multiple pollutants. In Fig. 5, a high value represents a high probability of hazardous zone and a high risk of soil pollution. As a result of industrial activities, the pollution of agricultural soils with heavy metals has become very serious throughout the study area. Heavy-metal pollution of soil is dominated by anthropogenic activities in the study area (Lin et al., 2002, 2010). For validation, the crisp multiple-pollutant hazardous zone in Fig. 5 (b) is highlighted in red. A hazardous zone is delineated if the average concentrations of 1000 SIS realizations (Fig. 3) are higher than the regulatory thresholds. The figure identifies the areas where the soil is polluted by heavy metals to facilitate remediation efforts. However, the delineation of a hazardous or non-hazardous zone is crisp and it is hard to account for the imprecision and uncertainty. The probability map of hazardous zones provides further insights for identifying pollution sources, high risk areas and pollution pathways for use in subsequent management projects, such as soil remediation, risk assessment and additional investigations (Lin et al., 2010).

3.3. Multiple-variable indicator semivariogram based on cLHS and SRS samples

Soil sampling for mapping possible areas of contamination is likely to be costly and time-consuming, especially in cases where laboratory procedures are expensive or the investigated area is large (Zhao et al., 2008). In this study, we investigate the relationships between the

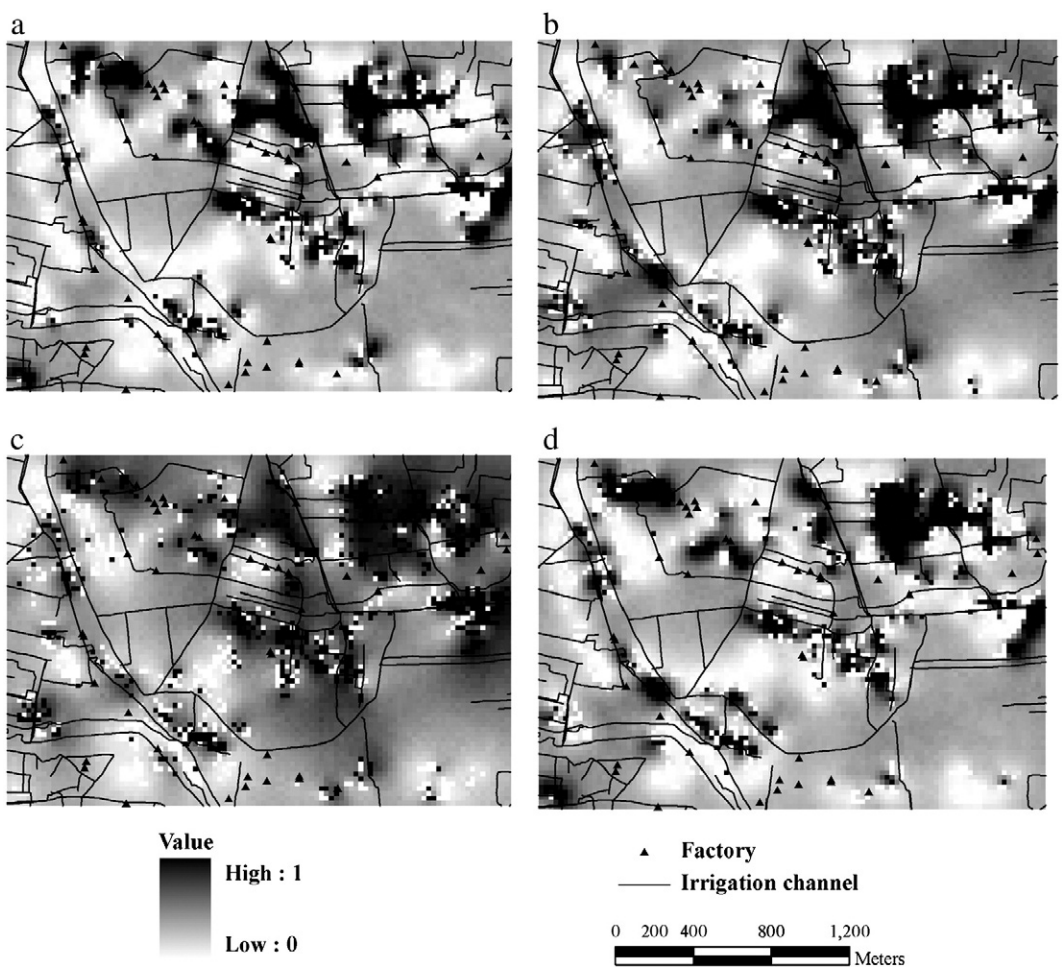

Fig. 4. The probability maps of (a) Cr, (b) Cu, (c) Ni, and (d) Zn in 1000 SIS realizations based on 1082 samples.

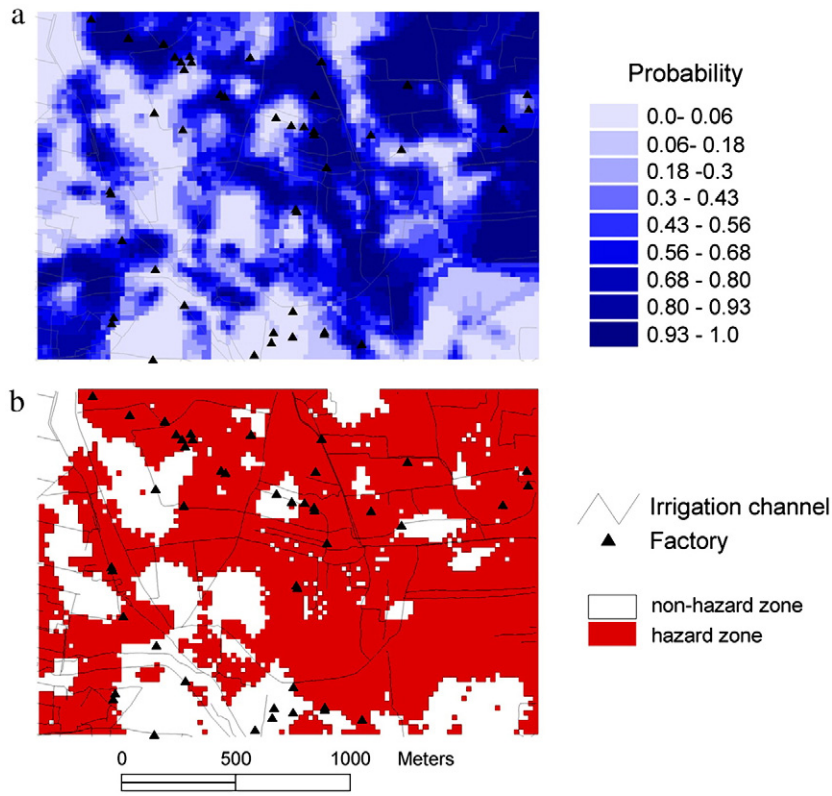

Fig. 5. (a) The probability map of hazard zone using MVIK based on 1082 samples and (b) crisp hazard delineation considering SIS average concentration maps.

sample patterns and hazard zone delineation. Table 2 details the indicator variogram analyses of multiple-variable indicators using various samples. The fitted models of the indicator variograms are the exponential models. Spatial structure analysis is widely regarded as a useful tool for illustrating the spatial patterns of variables. It is also a necessary basis for a number of other spatial analysis procedures, such as kriging analysis (Wang and Qi, 1998). Usually, the three most important features of a typical variogram model are the range, sill, and nugget effect (Lin et al., 2009). The sill, which is the upper limit that a variogram approaches at a large distance, is a measure of the variability of the investigated variable. The nugget effect is exhibited by the apparent non-zero value of the variogram at the origin, which may be due to the small-scale variability of the investigated process and/or measured errors. In the study, the range of an exponential variogram model is the distance lag at which the semivariance reaches 95% of the sill. It reveals the distance above which the variables become spatially independent. In indicator variograms based on cLHS samples, the fitted ranges, the nugget effects and the sills are 186–237, 0.0191–0.0312 and 0.246–0.236 m (Table 2), respectively. The results show that the sill increases and the range decreases as the sample size becomes smaller. On the other hand, the fitted ranges, the nugget effects and the sills are 225–9330, 0.0210–0.1761 and 0.234–0.443 m respectively in indicator variograms based on SRS samples (Table 2). The semivariance of 900 and 700 samples is similar to that of the 1082 samples. Based on the cases, the sill and the nugget effect increase when the sampling size decreases, which means a higher sill corresponds to greater variability in the variable. Moreover, patterns are smooth and probably related to the interrelated variation of the hazard zone in cases where the ranges of the variograms for 300 and 500 SRS samples are significantly large. These results exhibit a spatial correlation over a long distance under small SRS sample sizes. The experimental indicator variogram of 300 sub-samples misestimates the spatial patterns of the original samples. These results demonstrate that, as the number of samples decreased from 1082 to 300, the indicator variogram could not capture the spatial structures of the heavy-metal data. The results also show that the data configuration, such as the sample size and spatial configuration, affects the variogram (Wang and Qi, 1998; Jardim and Ribeiro, 2007; Juang et al., 2008). It is noteworthy that the cLHS results provide sufficient information to interpret the multiple-contaminant hazard delineation.

Table 2
Indicator variogram models for the MVIK based on cLHS and SRS samples.

Sample size	Method	Model	Parameters			RSS	r^2
			C0	C0 + C	R (m)		
1082		Exp.	0.0230	0.2368	237	2.16E-03	0.764
900	cLHS	Exp.	0.0226	0.2362	237	3.38E-03	0.711
700		Exp.	0.0191	0.2371	230	4.03E-03	0.701
500		Exp.	0.0238	0.2426	186	2.39E-03	0.721
300		Exp.	0.0312	0.2467	186	4.70E-03	0.703
900	SRS	Exp.	0.0209	0.2368	235	2.17E-03	0.797
700		Exp.	0.0210	0.2342	225	2.15E-03	0.714
500		Exp.	0.1501	0.3080	3135	2.42E-03	0.773
300		Exp.	0.1761	0.4432	9330	1.42E-03	0.794

Exp.: Exponential model; C0: Nugget; C0 + C: Sill; R: Range; RSS: Residual Sum of Squares; r^2: Coefficient of determination.

3.4. Comparison of the hazard probability under cLHS and SRS

Figs. 6 and 7 indicate that the MVIK probability of any soil contamination exceeds the corresponding regulation thresholds (z_k) based on cLHS and SRS sample patterns (i.e. 900, 700, 500, and 300 samples, respectively). MVIK estimates the uncertainty of unsampled

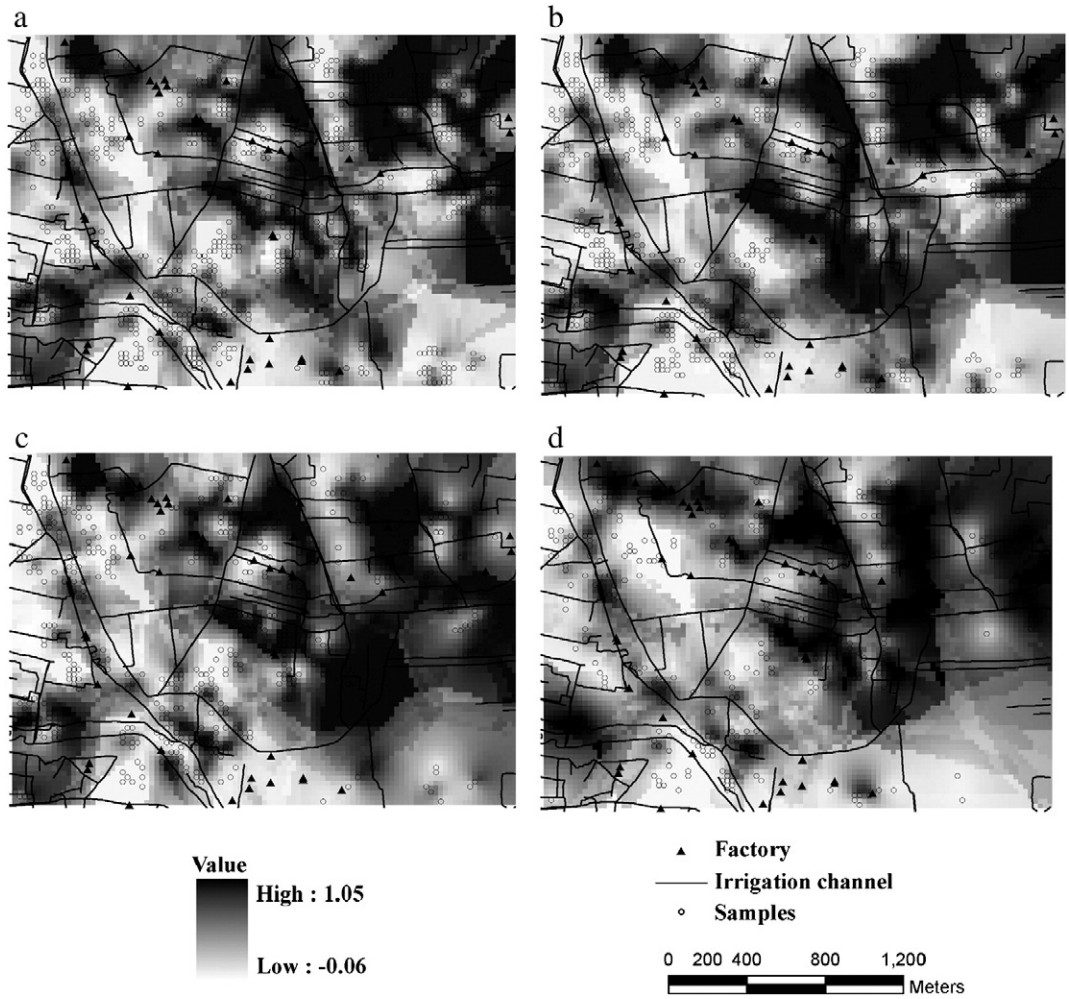

Fig. 6. The probability map of hazard zone using MVIK based on cLHS samples: (a) 900 samples, (b) 700 samples, (c) 500 samples, and (d) 300 samples.

values, which usually takes the form of a map of the probability that soil pollution exceeds the regulatory thresholds (Goovaerts, 1999). In Fig. 5 (a), the features of the hazardous zone probability based on the cLHS samples are better than those based on the SRS samples, especially when the data are sparse. The results indicate that the hazard probability map based on cLHS samples is consistent with that based on the original data. As the number of cLHS samples exceeds 500 (46.2% of 1082 samples), the distribution of hazard probabilities is similar to that based on 1082 samples. However, the patterns of hazard probabilities based on 500 and 300 SRS samples are fuzzy (Fig. 7 (c) and (d)). The results also indicate that the sampling density and pattern can have a significant effect on the result of hazard delineation. Table 3 shows the mean estimated variances and errors of the MVIK hazard probability map for various sample sizes and methods. The mean estimated variances using cLHS samples are between 0.1703 and 0.2252 mg/kg^2, while those using SRS samples are between 0.1706 and 0.2317 mg/kg^2. For the probability map based on 1082 samples (Fig. 5 (a)), the mean estimated variance is only 0.1702 mg/kg^2. Moreover, the mean estimated errors of the cLHS samples are 0.0549–0.1588 mg/kg, while those of the SRS samples are 0.0553–0.2182 mg/kg. It is also clear from the results that reducing the sample size could lead to estimation variances and errors rising. Comparison of the sampling methods shows that the estimated variances and errors of cLHS are lower than those of SRS. Using MVIK with cLHS samples is more effective in delineating hazardous zones than using MVIK with SRS samples. Furthermore, an effective sampling approach, such as cLHS, for multivariate distributions can be used to replicate the spatial structures and patterns of the investigated heavy metals.

In this study, we adopted MVIK hazard probabilities of 0.95, 0.85, 0.75, 0.5 and 0.25 to delineate safe and hazardous zones for soil pollutants and characterize the uncertainty based on various probabilities (Table 4). The results show that the areas with hazardous levels of heavy-metal pollutants (i.e., where the MVIK probability exceeds 0.85) cover over 20% of the study area. Moreover, the delineated area, where the probability is in excess of 0.25, is similar to that of the crisp hazard zone (i.e., the ratio of the hazardous area to the total is 69.7% in Fig. 5 (b)). The results also show that MVIK provides a means of identifying the hazardous zones. Table 4 also reports the proportions of hazardous areas to total areas for the cLHS and SRS cases. The proportions of hazardous area using cLHS samples are closer to the area proportions based on 1082 samples than the proportions derived by using SRS samples (Table 4). The findings show the cLHS approach is more reliable than the SRS approach, especially when the number of samples is small. The cLHS approach provides full coverage of each variable by maximally stratifying the distribution of the samples. Furthermore, safe and potentially hazardous regions containing heavy metals were delineated according to various probabilities estimated by MVIK. The probabilistic results of the classifications provide an alternative way to explore the spatial uncertainty of hazards and help government administrators establish a sound policy for the management of soil contamination.

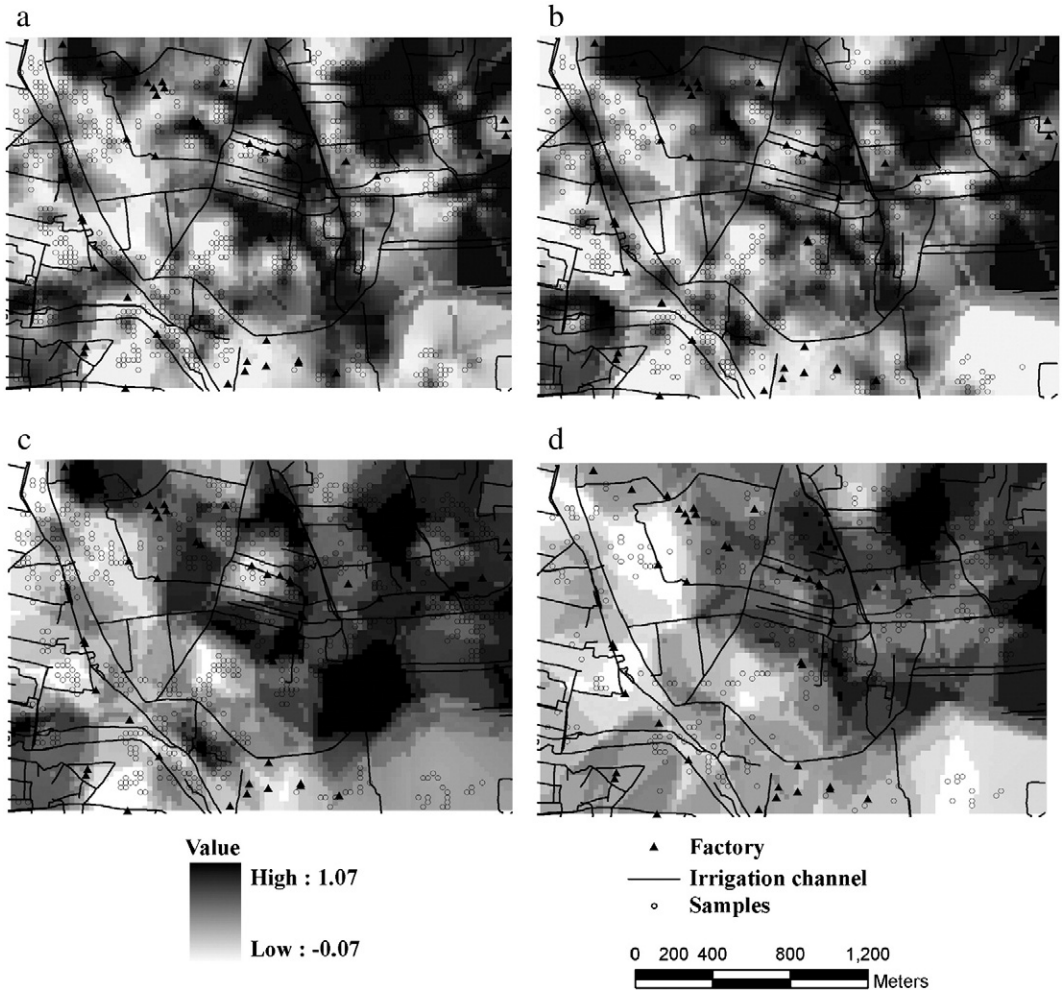

Fig. 7. The probability map of hazard zone using MVIK based on SRS samples: (a) 900 samples, (b) 700 samples, (c) 500 samples, and (d) 300 samples.

4. Conclusion

The principle of environmental pollution monitoring is based on the variability and uncertainty of hazardous zones. This study utilizes multiple-variable indicator kriging (MVIK) with sufficient samples to delineate hazardous zones and quantify the risk of multiple pollutants in contaminated soil. The results demonstrate that MVIK based on cLHS samples is an alternative means of determining the hazard probability of pollutants and identifying the risk of hazard delineation. The estimation variances and errors are strongly related to the number and configuration of samples. The cLHS approach offers a reasonably efficient way to ensure good coverage of the data and to replicate the distribution of multiple pollutants when compared to random sampling. In the study area, the hazardous zones are in the vicinity of industrial plants and irrigation systems, so future sampling will be denser in areas that are deemed critical. Furthermore, application of proposed models can provide further insight into identifying hotspots and hazardous areas from complex field data.

Table 3
Mean estimated variances and errors of the MVIK based on cLHS and SRS samples.

	Mean estimated variances $(mg/kg)^2$		Mean estimated errors (mg/kg)	
Sample size	cLHS	SRS	cLHS	SRS
900	0.1703	0.1706	0.0549	0.0553
700	0.1743	0.1754	0.0901	0.1234
500	0.1999	0.2209	0.0970	0.1613
300	0.2252	0.2317	0.1588	0.2182

Acknowledgment

The authors wish to thank the anonymous reviewers for their valuable comments and suggestions.

Table 4
Ratio of hazardous area to total area in various samples for five threshold levels.

	Sample size				
		900	700	500	300
	Method				(Unit: %)
Pr $ob[Z(x)>0.95\|(n)]$	cLHS	15.0	15.7	14.9	9.1
	SRS	14.5	14.6	8.8	3.0
Pr $ob[Z(x)>0.85\|(n)]$	cLHS	22.1	21.9	23.6	21.9
	SRS	21.3	22.3	24.4	12.4
Pr $ob[Z(x)>0.75\|(n)]$	cLHS	28.8	30.6	33.8	30.3
	SRS	27.8	31.6	33.1	18.0
Pr $ob[Z(x)>0.50\|(n)]$	cLHS	49.3	52.0	58.4	51.4
	SRS	48.8	53.4	58.3	42.2
Pr $ob[Z(x)>0.25\|(n)]$	cLHS	69.9	70.2	77.0	75.7
	SRS	69.9	73.6	84.3	78.2

References

Amini, M., 2005. Mapping risk of cadmium and lead contamination to human health in soils of Central Iran. Science of the Total Environment 347 (1–3), 64–77.

Carre, F., McBratney, A.B., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141 (1–2), 1–14.

Deutsch, C.V., Journel, A.G., 1992. GSLIB. Geostatistical Software Library and User's Guide. Oxford University Press, New York. 340 pp.

Diodato, N., Ceccarelli, M., 2004. Multivariate indicator kriging approach using a GIS to classify soil degradation for Mediterranean agricultural lands. Ecological Indicators 4 (3), 177–187.

Franco, C., Soares, A., Delgado, J., 2006. Geostatistical modelling of heavy metal contamination in the topsoil of Guadiamar river margins (S Spain) using a stochastic simulation technique. Geoderma 136 (3–4), 852–864.

Goovaerts, P., 1997. Kriging vs. stochastic simulation for risk analysis in soil contamination. In: Soares, A., Go´mez-Hern´andez, J., Froidevaux, R. (Eds.), geoENV I—Geostatistics for Environmental Applications. Kluwer Academic Publishers, Dordrecht, pp. 247–258.

Goovaerts, P., 1998. Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. Biology and Fertility of soils 27 (4), 315–334.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89 (1–2), 1–45.

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1–2), 3–26.

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 (1), 31–48.

GS+, 2004. Geostatistics for the Environmental Sciences. Version 5.0. Gamma Design Software, Plainwell, MI.

Halvorson, J.J., Smith, J.L., Papendick, R.I., 1996. Integration of multiple soil parameters to evaluate soil quality: a field example. Biology and Fertility of Soils 21 (3), 207–214.

Hassan, M.M., Atkins, P.J., 2007. Arsenic risk mapping in Bangladesh: a simulation technique of cokriging estimation from regional count data. Journal of Environmental Science and Health, Part A Toxic/Hazardous Substances and Environmental Engineering 42 (12), 1719–1728.

Jang, C.S., Chen, S.K., Ching-Chieh, L., 2008. Using multiple-variable indicator kriging to assess groundwater quality for irrigation in the aquifers of the Choushui River alluvial fan. Hydrological Processes 22 (2), 4477–4489.

Jardim, E., Ribeiro, P.J., 2007. Geostatistical assessment of sampling designs for Portuguese bottom trawl surveys. Fisheries Research 85 (3), 239–247.

Juang, K.W., Chen, Y.S., Lee, D.Y., 2004. Using sequential indicator simulation to assess the uncertainty of delineating heavy-metal contaminated soils. Environmental Pollution 127 (2), 229–238.

Juang, K.W., Liao, W.J., Liu, T.L., Tsui, L., Lee, D.Y., 2008. Additional sampling based on regulation threshold and kriging variance to reduce the probability of false delineation in a contaminated site. Science of the Total Environment 389 (1), 20–28.

Kasassi, A., Rakimbei, P., Karagiannidis, A., Zabaniotou, A., Tsiouvaras, K., Nastis, A., Tzafeiropoulou, K., 2008. Soil contamination by heavy metals: measurements from a closed unlined landfill. Bioresource Technology 99 (18), 8578–8584.

Krishna, A.K., Govil, P.K., 2004. Heavy metal contamination of soil around Pali industrial area, Rajasthan, India. Environmental Geology 47 (1), 38–44.

Lee, J.J., Liu, C.W., Jang, C.S., et al., 2008. Zonal management of multi-purpose use of water from arsenic-affected aquifers by using a multi-variable indicator kriging approach. Journal of Hydrology 359 (3–4), 260–273.

Lin, Y.P., 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 (8), 900–909.

Lin, Y.P., 2008. Simulating spatial distributions, variability and uncertainty of soil arsenic by geostatistical simulations in Geographic information systems. Open Environmental Sciences 2, 26–33.

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 (1), 189–199.

Lin, Y.P., Teng, T.P., Chang, T.K., 2002. Multivariate analysis of soil heavy metal pollution and landscape pattern in Changhua County in Taiwan. Landscape and Urban Planning 62 (1), 19–35.

Lin, Y.P., Chu, H.J., Wang, C.L., Yu, H.H., Wang, Y.C., 2009. Using remote sensing data with Latin hypercube sampling and geostatistical approach in delineating and sampling landscape changes induced by chronological large-extent physical disturbances. Sensors 9 (1), 148–174.

Lin, Y.P., Cheng, B.Y., Shyu, G.S., Chang, T.K., 2010. Combining a finite mixture distribution model with indicator kriging to delineate and map the spatial patterns of soil heavy metal pollution in Chunghua County, central Taiwan. Environmental Pollution 158 (1), 235–244.

Lin, Y.P., Chu, H.J., Hwang, Y.L., Cheng, B.Y., Chang, T.K., accepted for publication. Modeling spatial uncertainty of heavy metals content in soil by conditional Latin hypercube sampling and geostatistical simulation. Environmental Earth Sciences.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1092.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computational Geoscience 32 (9), 1378–1388.

Norra, S., Lanka-Panditha, M., Kramar, U., Stuben, D., 2006. Mineralogical and geochemical patterns of urban surface soils, the example of Pforzheim, Germany. Applied Geochemistry 21, 2064–2081.

Oyedele, D.J., Amusan, A.A., Obi, A.O., 1996. The use of multiple-variable indicator kriging technique for assessment of the suitability of an acid soil for maize. Tropical Agriculture 73 (4), 259–263.

Post, J., Hattermann, F.F., Krysanova, V., Suckow, F., 2008. Parameter and input data uncertainty estimation for the assessment of long-term soil organic carbon dynamics. Environmental Modelling & Software 23, 125–138.

Schnabel, U., 2004. Uncertainty assessment for management of soil contaminants with sparse data. Environmental Management 33 (6), 911–925.

Smith, J.L., Halvorson, J.J., Papendick, R.I., 1993. Using multiple-variable indicator kriging for evaluating soil quality. Soil Science Society of America Journal 57, 743–749.

Van Meirvenne, M., Goovaerts, P., 2001. Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. Geoderma 102 (1–2), 75–100.

Wang, X.J., Qi, F., 1998. The effects of sampling design on spatial structure analysis of contaminated soil. Science of the Total Environment 224 (1–3), 29–41.

Xu, C.G., He, H.S., Hu, Y.M., Chang, Y., Li, X.Z., Bu, R.C., 2005. Latin hypercube sampling and geostatistical modeling of spatial uncertainty in a spatially explicit forest landscape model simulation. Ecological Modelling 185 (2–4), 255–269.

Zhang, C., Wu, L., Luo, Y., Zhang, H., Christie, P., 2008. Identifying sources of soil inorganic pollutants on a regional scale using a multivariate statistical approach: role of pollutant migration and soil physicochemical properties. Environmental Pollution 151, 470–476.

Zhao, Y.C., Xu, X.H., Huang, B., Sun, W.X., Shao, X.X., Shi, X.Z., Ruan, X.L., 2007. Using robust kriging and sequential Gaussian simulation to delineate the copper- and lead-contaminated areas of a rapidly industrialized city in Yangtze River Delta, China. Environmental Geology 52 (7), 1423–1433.

Zhao, Y., Xu, X., Sun, W., Huang, B., Darilek, J.L., Shi, X., 2008. Uncertainty assessment of mapping mercury contaminated soils of a rapidly industrializing city in the Yangtze River Delta of China using sequential indicator co-simulation. Environmental Monitoring and Assessment 138 (1–3), 343–355.