Environmental Pollution 158 (2010) 235–244

Contents lists available at ScienceDirect

Environmental Pollution

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

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

Lin Yu-Pin[a,*], Cheng Bai-You[a], Shyu Guey-Shin[b], Chang Tsun-Kuo[a]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, 1, Section 4, Roosevelt Road, Da-an District, Taipei City 106, Taiwan, ROC
[b] Department of Environmental Management, Tungnan University, 152, Section 3, PeiShen Road, ShenKeng, Taipei 222, Taiwan, ROC

Effectively determine pollution threshold and map contaminated areas.

A R T I C L E I N F O

Article history:
Received 10 December 2008
Received in revised form
14 July 2009
Accepted 17 July 2009

Keywords:
Finite mixture distribution model
Heavy metal pollution
Contamination threshold
Multivariate indicator kriging
Geographical information system

A B S T R A C T

This study identifies the natural background, anthropogenic background and distribution of contamination caused by heavy metal pollutants in soil in Chunghua County of central Taiwan by using a finite mixture distribution model (FMDM). The probabilities of contaminated area distribution are mapped using single-variable indicator kriging and multiple-variable indicator kriging (MVIK) with the FMDM cut-off values and regulation thresholds for heavy metals. FMDM results indicate that Cr, Cu, Ni and Zn can be individually fitted by a mixture model representing the background and contamination distributions of the four metals in soil. The FMDM cut-off values for contamination caused by the metals are close to the regulation thresholds, except for the cut-off value of Zn. The receiver operating characteristic (ROC) curve validates that indicator kriging and MVIK with FMDM cut-off values can reliably delineate heavy metals contamination, particularly for areas lacking background information and high heavy metal concentrations in soil.

1. Introduction

Data on heavy metals in soil occasionally exhibit complex spatial variations and patterns, possibly complicating efforts to identify the characteristics of heavy metal pollutants in soil, particularly in urban and industrial areas. Such complex variations and patterns are attributed to natural phenomena (natural background) and human activities (anthropogenic background) (Martín et al., 2006; Micó et al., 2006; El Sebai et al., 2007; Lin, 2008). Anthropogenic activities, such as the discharge of waste from metal processing plants, burning fossil fuels and pesticides use, subsequently caused most soil pollution (Lin et al., 2002a). Some of these anthropogenic background-related practices help distinguish the anthropogenic background from a natural background (Portier, 2001). Other issues are related to how data about anthropogenic background concentrations may be addressed in policy decisions related to pollution remediation (Portier, 2001), risk assessment, and environmental management in investigated sites. Thus, determining precise pollution thresholds depends on the ability to develop efficient analytical approaches for identifying natural and anthropogenic background concentrations of heavy metals in soil. Such thresholds can then be used to delineate and map polluted areas, as well as determine the spatial distributions and variability of heavy metals in soil, for environmental policy decision makers.

The concentrations of various elements, including heavy metals in soil, provide a valuable reference for establishing threshold values or action levels for screening study decisions (Portier, 2001). Soil pollution studies normally take samples from mixtures of polluted and unpolluted areas without understanding the full areal extent of contamination. Therefore, the soil samples contain a mixture including natural and anthropogenic concentrations of heavy metals in soil. A finite mixture model assumes that the samples taken from a mixture distribution are composed by several component distributions of an investigated variable in an area of interest. Meanwhile, finite mixtures of distributions can be used in mathematical approaches for statistical modeling of an extensive variety of random phenomena (McLachlan and Peel, 2000). Therefore, a mixture model of soil samples can model complex distributions by choosing the components, i.e., natural and anthropogenic concentrations, that represent local areas in the actual distributions of soil concentrations. Soil pollution studies can conceptualize and model the overall background concentration distribution as a mixture of a natural background distribution, an anthropogenic background distribution and a distribution designed

* Corresponding author. Tel.: +886 2 33663467; fax: +886 2 33663464.
E-mail address: yplin@ntu.edu.tw (Y.-P. Lin).

0269-7491/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envpol.2009.07.015

to accommodate the potential for contamination site samples included in a background sample set (Portier, 2001).

Mapping polluted areas based on given pollution thresholds is essential because related information significantly contributes to efforts to formulate environmental action strategies, such as the remediation of soil polluted by heavy metals, after sampling the soil and determining pollution thresholds. In indicator kriging, assumptions are not made about the underlying invariant distribution, and 0–1 indicator transformations of data render the predictor robust to outliers (Cressie, 1993). Indicator kriging is widely used to map the probabilities of estimates that exceed given threshold levels such as regulatory thresholds of heavy metal contamination in soil (Van Meirvenne and Goovaerts, 2001; Lin et al., 2002a; Brus et al., 2002; Cattle et al., 2002; Juang et al., 2004; Schnabel et al., 2004; Zhao et al., 2007; Ungaro et al., 2008). Moreover, the soil in a given site may be polluted by several heavy metals, e.g., an agricultural area already urbanized and industrialized (Albanese et al., 2007; Norra et al., 2006; Wang, 2002; Huang et al., 2007; Lin et al., 2002b; Manta et al., 2002; Lin, 2008). Mapping areas polluted with multiple heavy metals requires using multivariate analysis with spatial estimation methods, such as factorial kriging and multivariate indicator kriging (Facchinelli et al., 2001; Lin et al., 2002a; Juang et al., 2004; Zhang et al., 2008; Rodríguez et al., 2008; Verstraete and Van Meirvenne, 2008). By utilizing the finite mixture distribution model (FMDM), this study identifies the natural background, anthropogenic background and distribution of contamination caused by heavy metal pollutants, with their respective cut-off values, in the soil in Chunghua County of central Taiwan. Additionally, the spatial probability of contaminated areas in the study area is mapped using indicator kriging and multiple-variable indicator kriging with the thresholds derived by the FMDM, as well as regulatory thresholds for heavy metals in soil. Moreover, indicator kriging and multiple indicator kriging with FMDM cut-off values and regulatory thresholds are validated using the ROC curve by classifying heavy metal contaminants in soil. Finally, contaminated areas and sources for subsequent remediation action are identified by inputting indicator kriging maps into a geographical information system (GIS) with land-use data.

2. Data and methods

2.1. Study area and sampling data

In this study, 1309 topsoil (0–15 cm) samples containing As, Cd, Cr, Cu, Hg, Ni, Pb and Zn collected between February and August 2002 were obtained from the soil heavy metal investigation project of the Environmental Protection Administration (EPA) of the Republic of China, Taiwan (Fig. 1) based on previous investigations in 1992–1999. Approximately 1 kg of each soil sample was collected using a stainless steel spade and a plastic scoop and, then, stored in a plastic food bag. The sampling location coordinates were recorded by GPS. The sampling location strategy is not uniform, which was based on the irregularity of farmland (approximately 461.0 ha) in the study area. Boundary of the study area is not a regular area caused by the topography of irrigation systems.

After air drying at room temperature, 3 g soil samples were disaggregated, sieved to 0.85 mm (20 mesh) and ground to a fine 0.15 mm (100 mesh) 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 Cd, Cr, Ni, Cu, Zn, and Pb in the samples were determined by Inductively Coupled Plasma-Optical Emission Spectrometers (ICP-OES). For Hg determination, 0.5 g samples were digested with 5 mL H_2SO_4 and 2 mL HNO_3 and, then, heated to 121 $\pm$ 3 °C for

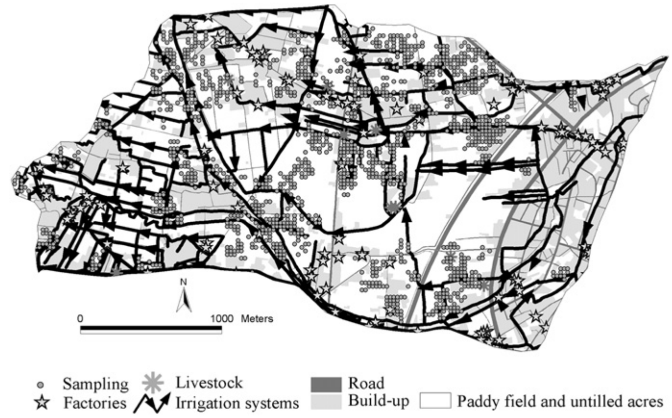

Fig. 1. Locations of samples in Chunghua County, central Taiwan.

15 min. The HG content was then determined by analyzing the extract using cold vapor absorption atomic spectrometry (CVAAS). Approximately 10 mL of H_2O_2 (30%) and 30 mL of 9.6 M HCl were added to 1 g of the air-dried sample. Next, arsine was generated in a continuous flow system, and $NaBH_4$ solution was used as the carrier solution. Finally, the sample was diluted with water in a 50 mL volumetric flask, followed by determination of the As content using Hydride Generation Atomic Absorption Spectroscopy (HGAAS).

2.2. Finite mixture distribution model

Let $Z_1, \ldots, Z_n$ denote a random sample of size n, where Z_i represents a p-dimensional random vector with the probability density function $f(x_i)$ on R^p (McLachlan and Peel, 2000). The probability density function of the random vectors Z_i under a k-component mixture model is written in parametric form (Titterington et al., 1985) as follows:

$$f(z) = \sum_{i=1}^{k} \pi_i f_i(z) = \pi_1 f_1(z) + \cdots + \pi_k f_k(z) \qquad \sum_{i=1}^{k} \pi_i = 1, \quad (0 \le \pi_i \le 1) \tag{1}$$

where π_i denotes the mixed proportions or weights of $f_i(z)$.

In the above equation, $f_i(z)$ can be any distribution comprised of natural, anthropogenic and contamination distributions in heavy metal soil samples. In this study, a lognormal distribution is used, as shown in Eq. (2).

$$f_{LN(\mu,\sigma^2)}(z) = \frac{1}{\sqrt{2\pi}\sigma z} e^{-\frac{(\ln z-\mu)^2}{2\sigma^2}}, \quad x > 0 \tag{2}$$

The mixed distribution composed by lognormal distributions can be written as

$$f(z|\pi,\mu,\sigma) = \sum_{i=1}^{k} \pi_i f_i(z) = \pi_1 f_1(z|\mu_1,\sigma_1) + \cdots + \pi_k f_k(z|\mu_k,\sigma_k) \tag{3}$$

where μ_i and σ_i are the mean and standard deviation of $f_i(z)$, respectively.

Finally, the model parameters (π_i, μ_i, σ_i) are estimated using the maximum likelihood. The x^2 test is then performed to test hypothesis H_0: the heavy metal data in a soil sample follows a specified statistical distribution:

$$x^2 = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \tag{4}$$

and x^2 has $(m - k - 1)$ degrees of freedom, where m denotes the number of classes of the original population and k is the number of estimated parameters (Krebs, 1999; Liu et al., 2002). Here, the maximum likelihood approach is implemented using Rmix software for estimating parameters of distributions and for applying the x^2 test to the model (the overall distribution) (Du, 2002). The Rmix software is an implementation of the MIX software (MacDonald and Pitcher, 1979) ported into R software (Ihaka and Gentleman, 1996). The cut-off value (z') should be satisfied by

$$\pi_i f_i(z'|\mu_i, \sigma_i) = \pi_{i+1} f_{i+1}(z'|\mu_{i+1}, \sigma_{i+1}) \tag{5}$$

2.3. Determination of threshold values for heavy metal concentrations in soil

Pollution threshold values are determined by applying the FMDM concepts of Portier (2001) to the data on heavy metals in soil. In the first mixed distribution (i.e., natural background and contaminated), the population probability density function (PPDF) becomes a mixture of observations of background areas and contaminated areas (Fig. 2a). Sample concentrations from contaminated areas have an increased probability for high concentration values (Fig. 2a). The distribution of mixing background and anthropogenic distributions shifts the mean and median values to the right of the background levels and increases the underlying spread or variance of the distribution. A situation in which the background PPDF for one soil type significantly differs from that of an adjacent soil type the overall PPDF will be the second mixed distribution (i.e., background, anthropogenic and contaminated) (Fig. 2b). Anthropogenic concentration close to background concentration could be caused by human activities, but not related to contamination activities such as waste discharge from metal processing plants, fossil fuel burning and use of pesticides. The third mixed distribution comprises a natural background mixed with low and high contaminated samples (Fig. 2c). Low and high contaminated concentrations could be caused from contamination activities and serious contamination activities, e.g., highly polluted water used to irrigate paddy fields. Samples from low and high contaminated areas significantly differ in concentrations and probabilities in PPDF (Fig. 2c). In this study, the cut-off values (Fig. 2d) are determined by following equation (Titterington et al., 1985):

$$\pi_i \int_{x_0}^{+\infty} f_i(z)dz = \pi_{i+1} \int_{-\infty}^{x_0} f_{i+1}(z)dz \tag{6}$$

2.4. Indicator kriging and multivariate indicator kriging

The probabilities in a square (25 m × 25 m) grid comprised of 1309 cells transferred to the GIS system for display are estimated using Indicator kriging and MVIK. Indicator kriging estimates the probability of exceeding specific threshold values, z_k, at a given location (Lin et al., 2002a). In indicator kriging, the data, $z(x)$, is transformed into an indicator as follows:

$$i(x, z_k) = \begin{cases} 1, & \text{if } z(x) \le z_k \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

At an unsampled location, x_0, the probability $z(x) \le z_k$ can be estimated using a linear combination of neighboring indicator variables. This ordinary indicator kriging estimator is

$$\mathrm{Prob}[z(x_0) \le z_k/(n)]^* = \sum_{\alpha=1}^{n} \lambda_\alpha i(x_\alpha; z_k) \tag{8}$$

where $i(x_\alpha; z_k)$ denotes the indicator values at x_α; $\alpha = 1, \cdots, n$; and λ_α represents the kriging weight of $i(x_\alpha; z_k)$ determined by solving the following kriging system when estimating $\mathrm{Prob}[z(x_0)] \le z_k/n]$.

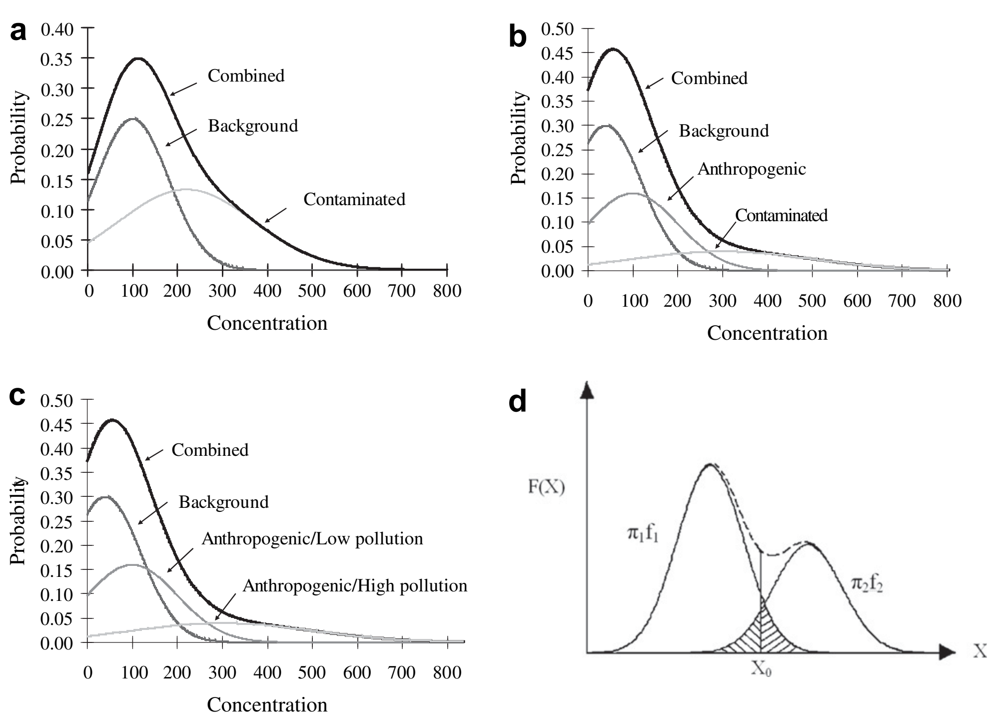

Fig. 2. Definitions of (a) the first type of finite mixture distribution model (redraw from Portier, 2001), (b) the second type of finite mixture distribution model (redraw from Portier, 2001), (c) the third type of finite mixture distribution model, (d) the cut-off values finite mixture distribution model.

An ordinary indicator kriging system can be solved using by

$$\sum_{\beta=1}^{n} \lambda_\beta \gamma_i(x_\alpha - x_\beta; z_k) + \mu = \gamma_i(x_\alpha - x_0; z_k) \tag{9}$$

$$\sum_{\beta=1}^{n} \lambda_\beta = 1 \tag{10}$$

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

Multiple-variable indicator transformation (MVIT) (Smith et al., 1993; Diodato and Ceccarelli, 2004) considers the conditional cumulative distribution function (ccdf) of variables (Cu, Cr, Ni and Zn) as follows:

$$F_1(x; z_1|(n_1)) = \mathrm{prob}\{Cr(x) > z_1(n_1)\} \tag{11}$$

where (n_1) denotes the conditioning to $n_1$data, while z_1 denotes the threshold for Cr.

$$F_2(x; z_2|(n_2)) = \mathrm{prob}\{Cu(x) > z_2|(n_2)\} \tag{12}$$

with n_2 data and the z_2 threshold for Cu.

$$F_3(x; z_3|(n_3)) = \mathrm{prob}\{Ni(x) > z_3|(n_3)\} \tag{13}$$

with n_3 data and the z_3 threshold for Ni.

$$F_4(x; z_4|(n_4)) = \mathrm{prob}\{Zn(x) > z_4|(n_4)\} \tag{14}$$

with n_4 data and the z_4 threshold for *Zn*.

At an unsampled point x, F_i is estimated as a linear combination of sample data falling within a neighborhood of x using ordinary kriging:

$$F_i(x; z_i|(n_i)) = E\{I_i(x; z_i)|(n_i)\} \tag{15}$$

where $n(x)$ denotes the number of neighboring sample points x; and $\lambda_{\alpha i}$ represents the weight factors depending on parameter (i) and sample point (α) calculated by the kriging systems.

$$[F_i(x; z_i|(n_i))]^*_{\mathrm{OK}} = \sum_{\alpha=1}^{n(x)} \lambda_{\alpha i}(x; z_i) I_i(x_\alpha; z_i) \tag{16}$$

2.5. Validation of soil contaminations

By using the Receiver Operating Characteristic (ROC) curve, this study validates the delineation and classification of contamination using indicator and multiple-variable indicator kriging with FMDM cut-off values and regulatory thresholds. The ROC curve is constructed by calculating the sensitivity and specificity of the resulting classification for each possible classification of contaminations as follows (Alatorre and Beguería, 2009):

$$\mathrm{senitiviity} = \frac{a}{a+c} \tag{17}$$

$$\mathrm{specificity} = \frac{b}{b+d} \tag{18}$$

where a and d, respectively, are true positives and true negatives, and b and c the false positives (type I error) and false negatives (type II error), respectively. The ROC value, i.e., a scalar measure of forecast discrimination, refers to the area under the curve that links the proportions of true positives *versus* the proportion of false positives for an infinite number of cut-off values.

3. Results and discussion

3.1. Finite mixture distribution models of heavy metal contaminants in soil

Table 1 summarizes both the basic statistics of heavy metals in soil in Taiwan and the EPA regulatory thresholds for those metals. The measured concentrations of Cr, Cu, Ni, and Zn, exceed the regulatory thresholds in 286 samples (21.8% of all samples) of Cr, 395 samples (30.2% of all samples) of Cu, 622 samples (47.5% of all samples) of Ni and 336 samples (25.7% of all samples) of Zn. Only 26 samples have Cd concentrations and one sample has a Hg concentration exceeding the regulatory thresholds. Notably, Cr, Cu, Ni and Zn are the major pollutants in the study area, correlating with the results of Lin et al. (2002a,b).

The mixture distributions of four heavy metals, i.e., Cu, Cr, Ni and Zn, are fitted by using lognormal distributions (Table 2 and Fig. 3). Importantly, FMDM can identify background and anthropogenic background levels, as well as contamination concentrations of heavy metals from soil samples taken from mixtures of polluted and unpolluted area without knowing the full areal extent of the contamination. According to the finite mixture distribution results, the distributions of Cr and Cu are composed of two lognormal distributions with their statistical distribution parameters π and μ at the 0.05 significance level (p-value > 0.05; reject H_0: the heavy metal data in a soil sample follows a specified statistical distribution). The two-distribution mixture distribution consists of two lognormal distributions that represent the background distribution and contamination distribution of Cr and Cu in soil in the study area (Fig. 3). Additionally, the distributions of Ni and Zn can consist of three lognormal distributions with parameters π and μ at the 0.05 significance level. The three-distribution mixture model comprises three lognormal distributions, which represent the background distribution, moderate contamination distribution and high contamination distribution of Ni and Zn in soil in the study area (Fig. 3). The fitted mixture distribution results indicate that, at the time of the study, the soil in the study area was contaminated by Cr, Cu, Ni and Zn from multiple sources, such as background concentrations and contamination (anthropogenic) concentrations. These mixture modeling results correspond to the indicator kriging results of Juang et al. (2001), Lin et al. (2002a,b), Yang and Chang (2005), and Verstraete and Van Meirvenne (2008), indicating that heavy metal data obeys lognormal distributions in the same study area. Above FMDM results further demonstrate that soil samples are taken from mixtures of low (moderate) polluted, high polluted and unpolluted areas across the entire study area.

3.2. Background concentrations of heavy metals in soil

FMDM results indicate that the geometric means of the background lognormal distributions of Cr, Cu, Ni and Zn are 97.2 mg kg^{-1}, 85.9 mg kg^{-1}, 108.4 mg kg^{-1} and 215.5 mg kg^{-1}, respectively (Fig. 3 and Table 2). Background concentrations of heavy metals in soil heavily depend on natural parent materials (Alloway, 1995; Kiekens, 1995; McGrath, 1995; El Sebai et al., 2007; Chen et al., 2009). Background concentrations of heavy metals in soil have been identified using the baseline approach (Chen et al., 1999; Dudka, 1993; Albanese et al., 2007) and multivariate analysis (Folkes et al., 2001; Martín et al., 2006; Martínez et al., 2007) to determine thresholds of pollutions. Alternatively, the backgrounds of heavy metals in soil have been determined using a basic statistical range, and the average or geometric mean of heavy metals in soil samples

Table 1
Statistics of measured soil heavy metals.

	DL (mg kg^{-1})	Samples with value < DL	Minimum (mg kg^{-1})	Median (mg kg^{-1})	Maximum (mg kg^{-1})	Mean (mg kg^{-1})	Standard deviation	RT (mg kg^{-1})	Sample with value > RT
As	0.190	0	4.6	10.7	37.4	11.4	3.8	60	0
Cd	0.116	203	<0.116	0.93	18.0	1.34	1.46	5	26
Cr	2.98	0	22.6	119.0	3070.0	194.0	212.5	250	286
Cu	3.61	0	11.0	116.0	3810.0	194.7	222.7	200	395
Hg	0.038	12	<0.038	0.35	5.35	0.43	0.38	5	1
Pb	3.65	0	6.1	36.9	148.27	42.7	17.5	500	0
Ni	3.64	0	21.3	189.2	4020.0	271.3	259.0	200	622
Zn	5.88	0	60.5	337.0	7850.0	526.4	549.6	600	336

DL: detection limit; RT: regulation threshold.

(Alloway, 1995; Kiekens, 1995; McGrath, 1995; Chen, 2000). Chen (2000) found that the ranges of Cr, Cu, Ni and Zn in their 100 unpolluted rural soil samples (surface soil and subsoil; 0–100 cm) obtained from 787 ha throughout entire Taiwan were 22.9–98.9 mg kg^{-1}, 7.2–35.1 mg kg^{-1}, 18.6–66.7 mg kg^{-1} and 30.1–392.0 mg kg^{-1}, respectively. Moreover, the upper limits of the background total concentrations of Cr, Cu, Ni and Zn in the unpolluted samples were 100.0 mg kg^{-1}, 35.0 mg kg^{-1}, 60.0 mg kg^{-1} and 120.0 mg kg^{-1}, respectively. The geometric mean of the Cr background distribution is close to the upper limit of the Cr background reported by Chen (2000). The geometric mean of the Zn background distribution is higher than the upper limit of the Zn background, but within the range of Zn soil samples taken by Chen (2000). Additionally, the geometric means of the Cu and Ni background distributions exceed the upper limits of Cu and Ni backgrounds reported by Chen (2000). Nevertheless, the geometric means of Cr, Cu, Ni and Zn background distributions and the background concentrations reported by Chen (2000) with FMDM results can be used as a benchmark with additional soil samples to identify background concentrations of these heavy metals in the study area accurately.

3.3. *Cut-off values for mixture distributions of heavy metal contamination in soil*

Table 3 lists the cut-off values for determining pollution concentrations of Cr, Cu, Ni and Zn in soil at significance levels based on type I error. Type I error for pollution is defined as an incorrect rejection of the null hypothesis for contaminated soil. The cut-off values of Cr and Cu pollution are 260.0 mg kg^{-1} and 192.8 mg kg^{-1}, respectively. The values are close to the regulatory thresholds for Cr and Cu concentrations in soil. Additionally, the cut-off value (193.2 mg kg^{-1}) of the second distribution, defined as moderate pollution, since Ni in soil is close to the regulatory threshold (200 mg kg^{-1}). The cut-off values of moderate Ni and Zn pollution are 193.2 mg kg^{-1} and 353.5 mg kg^{-1}, respectively. Moreover, areas with Ni and Zn concentrations that exceed 290.2 mg kg^{-1} and 963.0 mg kg^{-1}, respectively, are considered highly contaminated. The regulatory thresholds for Cr, Cu and Ni in soil have roughly the same concentrations as the FMDM cut-off values for Cr, Cu and Ni, except for the regulatory threshold for Zn. The regulatory threshold (600 mg kg^{-1}) for Zn is significantly greater than the cut-off value (353.5 mg kg^{-1}) for moderate Zn contamination, and is significantly lower than the cut-off value (963.0 mg kg^{-1}) for high Zn contamination. The cut-off value of moderate Zn contamination is close to the original regulatory threshold (300.0 mg kg^{-1}).

Spatial variability often results from a combination of intrinsic and extrinsic variations result from natural variation in soil and human activities (El Sebai et al., 2007). According to the FMDM results for the Cu and Cr concentrations, Cu and Cr concentrations with a two-mixture distribution provide adequate cut-off values for identifying pollution characteristics and contaminated sites in the study area (Fig. 3). The three-mixture distributions of Ni and Zn in soil provide insight into moderate and high contamination scenarios to assist decision makers in formulating policies for remediation and environmental actions. Although the FMDM cut-off values for Zn contamination differ from the regulatory thresholds, the moderate and high pollution cut-off values provide various alternatives for subsequent environmental action. The FMDM results provide cut-off values for delineating contamination to reevaluate the regulatory thresholds of heavy metal contaminations in soils, particularly for areas lacking background information and with high soil contamination. Moreover, according to the FMDM results of soil Cr, Cu, Ni and Zn, soils in the study area were highly contaminated by Ni and Zn through serious contamination activities, particularly Zn contaminants in soil.

Table 2
Mixture distribution goodness of fit and cut-off values calculated.

	Group	Pi (π)	Mu (μ)	Sigma (σ)	Cut-off values	χ^2	df	p-value
Cr	2	0.790	117.8	80.54	260.0	42.25	51	0.804
		0.210	469.4	278.32				
Cu	2	0.686	101.8	64.8	192.8	23.68	21	0.309
		0.314	389.2	247.8				
Ni	3	0.501	119.8	56.5	193.2	42.31	43	0.501
		0.209	232.9	53.12	290.2			
		0.290	566.3	264.75				
Zn	3	0.515	232.9	95.4	353.5	57.82	71	0.870
		0.354	572.6	234.6	963.0			
		0.131	1503.0	615.7				

Pi, Mu, sigma: parameters of lognormal distribution.
Null hypotheses, H_0: the heavy metal data in a soil sample follows a specified statistical distribution.
Reject H_0 if p-value $\leqq \alpha = 0.05$.

3.4. *Indicator variography based on cut-off values and regulatory thresholds*

A relatively consistent set of best-fit models with the lowest reduced sum of squares (RSS) and the highest r^2 values were generated using GS + software in order to fit variograms based on use of a least squares model. Table 4 lists the variography results for Cr, Cu, Ni and Zn based on the cut-off values for moderate and high pollution and the regulatory thresholds for those metals. The indicator variograms of Cr, Cu, Ni and Zn with high nugget effect ratios (>34.2%) represent high levels of small-scale variations or observation errors in the heavy metal data. These variations may be due to the complex distributions of heavy metal background and pollution sources, such as industrial plants and irrigation channels,

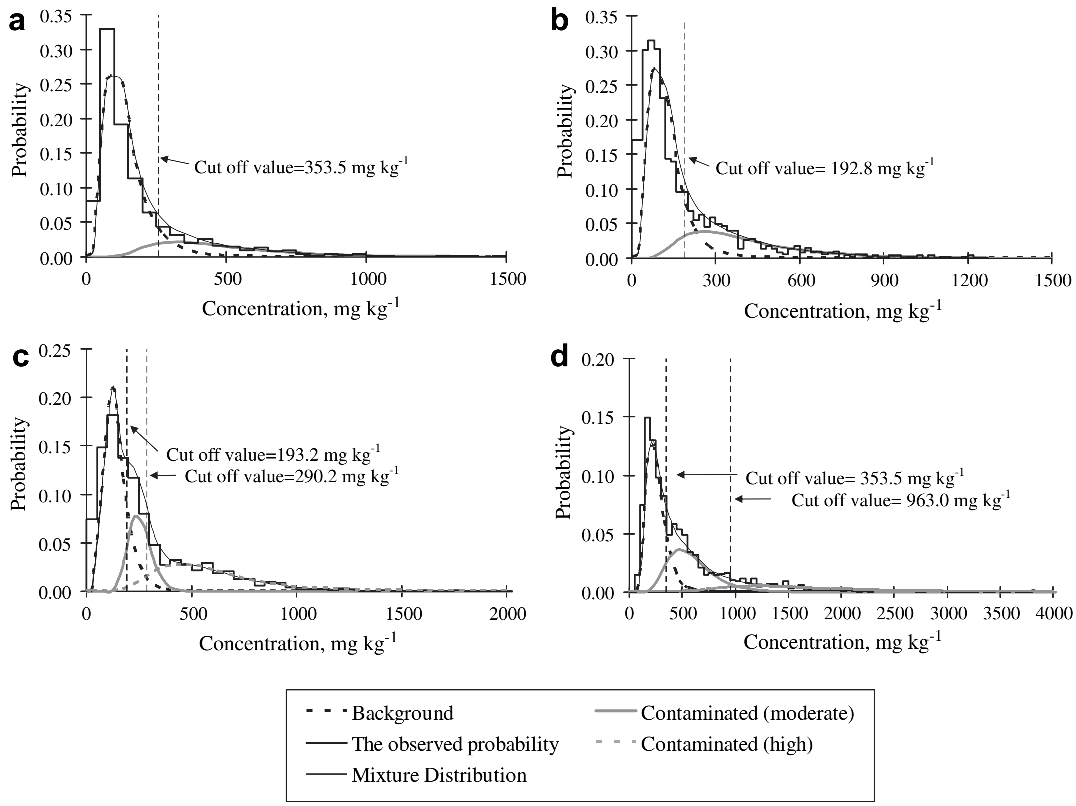

Fig. 3. The FMDM cut-off values for (a) Cr, (b) Cu, (c) Ni, (d) Zn in soil. (FMDM: finite mixture distribution model).

in the study area. Indicator variogram results further demonstrate that samples with Cr values exceeding the FMDM cut-off values for moderately contaminated soil with Cr and samples with Cr values exceeding regulatory thresholds have similar spatial contamination structures, i.e., indicator variogram models. The spatial structures of Cu and Ni in soil based on FMDM's moderate contamination cut-off values and the regulatory thresholds for Cu and Ni pollution are also similar.

Table 3
Cut-off values and RT used to determine pollution levels in soil based on type I error.

	FMDM			RT		
	Threshold Cut-off value (mg kg^{-1})	Distribution	Probabilities of type I error	Threshold RT (mg kg^{-1})	Distribution	Probabilities of type I error
Cr	260	F1[a]	0.04	250	F1[a]	0.05
	260	F2[a]	0.04	250	F2[a]	0.04
Cu	192.8	F1[b]	0.06	200	F1[b]	0.05
	192.8	F2[b]	0.06	200	F2[b]	0.06
Ni	193.2	F1[c]	0.05	200	F1[c]	0.04
	193.2	F2[c]	0.05	200	F2[c]	0.06
	290.2	F2[c]	0.03	200	F2[c]	0.06
	290.2	F3[c]	0.03	200	F3[c]	0.004
Zn	353.5	F1[d]	0.05	600	F1[d]	0.002
	353.5	F2[d]	0.05	600	F2[d]	0.22
	963.0	F2[d]	0.02	600	F2[d]	0.22
	963.0	F3[d]	0.02	600	F3[d]	0.002

[a,b]Mixture distribution composed by F1 (low concentration) and F2 (high concentration).
[c,d]Mixture distribution composed by F1 (low concentration), F2 (moderate concentration) and F3 (high concentration).
FMDM: finite mixture distribution model.
RT: regulation threshold.

3.5. *Indicator kriging with contaminated cut-off values and regulatory thresholds*

Figs. 4 and 5 show indicator kriged probability maps based on the FMDM cut-off values and regulatory thresholds for Cr, Cu, Ni and Zn. Areas with a high probability of exceeding the regulatory thresholds and contamination cut-off values for Cr in soil have similar spatial pollution distributions, which correlate strongly with the locations of irrigation channels and industrial factories (Fig. 4a). Most areas north of the study area have a high probability of exceeding the regulatory thresholds and FMDM contamination

Table 4
Variography results for heavy metals.

	Threshold (mg kg^{-1})	Model type	C_0 (mg kg^{-1})^2	$C_0 + C$ (mg kg^{-1})^2	Range (m)	R^2	RSS
Cr	250[a]	Exponential	0.068	0.180	171.0	0.959	2.04E-04
	260[b]	Exponential	0.082	0.175	216.0	0.953	2.26E-04
Cu	192.8[b]	Exponential	0.095	0.221	207.0	0.948	4.36E-04
	200[a]	Exponential	0.093	0.216	198.0	0.953	3.54E-04
Ni	193.2[b]	Exponential	0.079	0.226	267.0	0.916	1.24E-03
	200[a]	Exponential	0.076	0.226	255.0	0.920	1.14E-03
	290.2[b]	Exponential	0.099	0.198	204.0	0.900	8.64E-04
Zn	353.5[b]	Spherical	0.105	0.243	181.0	0.944	7.48E-04
	600[a]	Spherical	0.097	0.198	252.0	0.961	3.83E-04
	963[b]	Exponential	0.023	0.124	138.0	0.886	3.39E-04

FMDM: finite mixture distribution model.
RSS: model reduced sum of square.
C_0: Nugget.
$C_0 + C$: Sill.
[a] The regulatory thresholds for Cr, Cu, Ni, Zn in soil in Taiwan.
[b] The FMDM cut-off values.

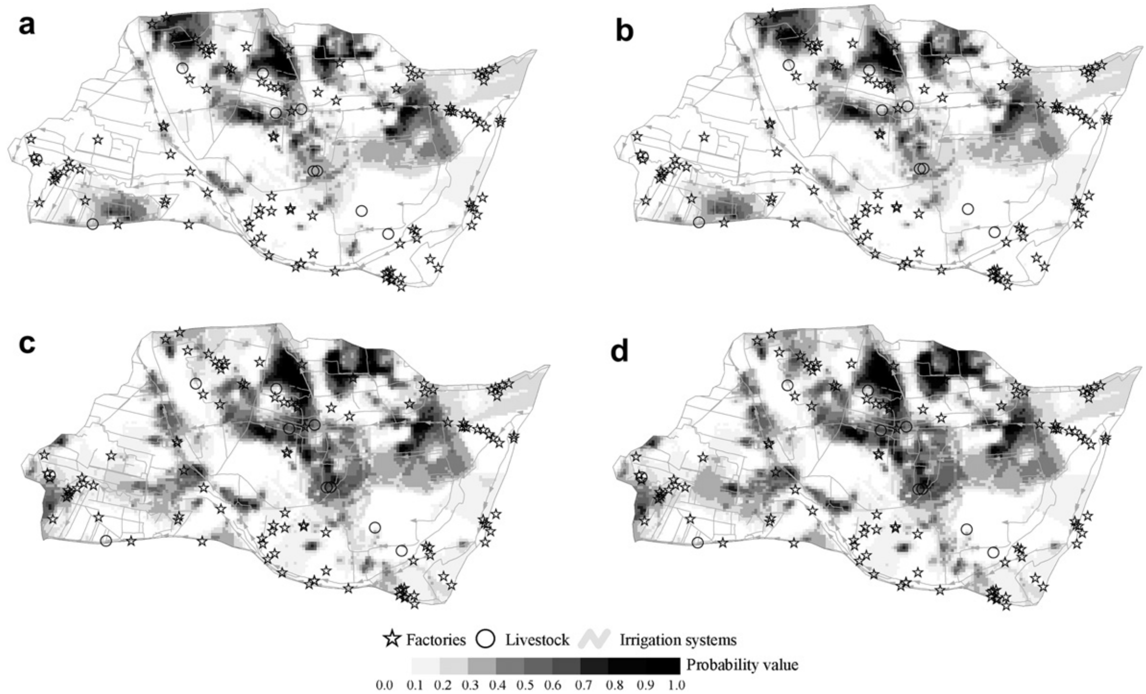

Fig. 4. Indicator kriged probability maps based on (a) regulatory thresholds of Cr = 250 mg kg^{-1}, (b) FMDM cut-off value Cr = 260.0 mg kg^{-1}, (c) regulatory thresholds of Cu = 200 mg kg^{-1}, (d) FMDM cut-off value Cu = 192.8 mg. (FMDM: finite mixture distribution model).

cut-off values (250 mg kg^{-1} and 260 mg kg^{-1}) for Cr in soil (Fig. 4a and b). The areas with high probabilities of exceeding the regulatory thresholds and contamination cut-off values (200 mg kg^{-1} and 192.8 mg kg^{-1}) for Cu in soil generally covered the central and north parts of the study area (Fig. 4c and d). The probability maps of areas exceeding both the regulatory thresholds and the FMDM cut-off values for Cr and Cu contamination indicate that soils were contaminated by wastewater from industrial factories, e.g., metal-working, electroplating and metal surface treatment. Most areas with a high probability of exceeding the regulatory thresholds and the FMDM cut-off values for Cu in soil were in the southern and northeaster locations of the study area, and were highly correlated with the locations of irrigation channels.

Based on the FMDM three-mixture distribution models for Ni and Zn, moderately and highly contaminated areas are also mapped (Fig. 5). Areas with a high probability of exceeding the regulatory threshold and moderate contamination cut-off value for Ni in soil have similar spatial distributions, which correlate with the locations of major irrigation channels and industrial factories in the study area, except for the western portion of the area (Fig. 5a). The probability map (Fig. 5c) of areas exceeding the high contamination cut-off value (290.2 mg kg^{-1}) indicates that highly contaminated areas are strongly correlated with the locations and patterns of irrigation channels and industrial factories in the northern part of the study area. Fig. 5d–f show the various patterns of Zn soil pollution based on the regulatory threshold (600 mg kg^{-1}), FMDM's moderate contamination cut-off values (353.5 mg kg^{-1}), and FMDM's high contamination cut-off values (963 mg kg^{-1}) for Zn in soil. Areas at a high risk of moderate Zn contamination are widely distributed over the study area (Fig. 5d). Areas with a high probability of Zn contamination correlated strongly with the locations of major irrigation channels in the study area, except for the southeastern parts of the study area (Fig. 5e). The sites with a high probability of exceeding the regulatory threshold for Zn contamination were distributed randomly throughout the northern parts of the study area (Fig. 5f). Although having their own characteristics when dispersed through different mediums, Cr, Cu, Ni and Zn in soil exhibit common spatial distributions that correlate with the locations of irrigation channels and industrial plants, especially in the central and northern parts of the study area.

The probability maps of Zn pollution in soil greatly facilitate local remediation and environmental action, such as conducting precise sampling and re-defining the local regulatory threshold for Zn in soil. Previous studies were only based on the regulatory thresholds for heavy metals in soil (Juang et al., 2001, 2004; Lin et al., 2002a). Conversely, the proposed probability maps of Cr, Cu, Ni and Zn exceeding the regulatory thresholds and FMDM cut-off values provide further insight into identifying pollution sources, high risk areas and pollution pathways for subsequent actions, such as soil remediation, additional investigations, risk assessments and local land-use planning.

3.6. Multiple indicator kriging with cut-off values and regulatory thresholds

Fig. 6a shows the conditional probability maps of sites that exceed the regulatory thresholds for Cr, Cu, Ni and Zn. According to those maps, sites at high risk of pollution by multiple heavy metals are strongly correlated with the locations of major irrigation channels in the northern parts of the study area. Fig. 6b shows the conditional probability maps of sites exceeding the cut-off values for Cr, Cu, moderate Ni and moderate Zn pollution. Comparing the conditional probability maps of sites that exceed the regulatory thresholds (Fig. 6a) and cut-off values (Fig. 6b) reveals that the northern part is at a high risk of soil contamination by multiple heavy metals. Meanwhile, the northeastern and southwestern parts should be deemed as moderately polluted areas with multiple heavy metals. Fig. 6c shows the probability maps of areas that exceed the regulatory thresholds for Cr, Cu, Ni and Zn in soil. Meanwhile, Fig. 6d shows the probability maps of areas exceeding the cut-off values for

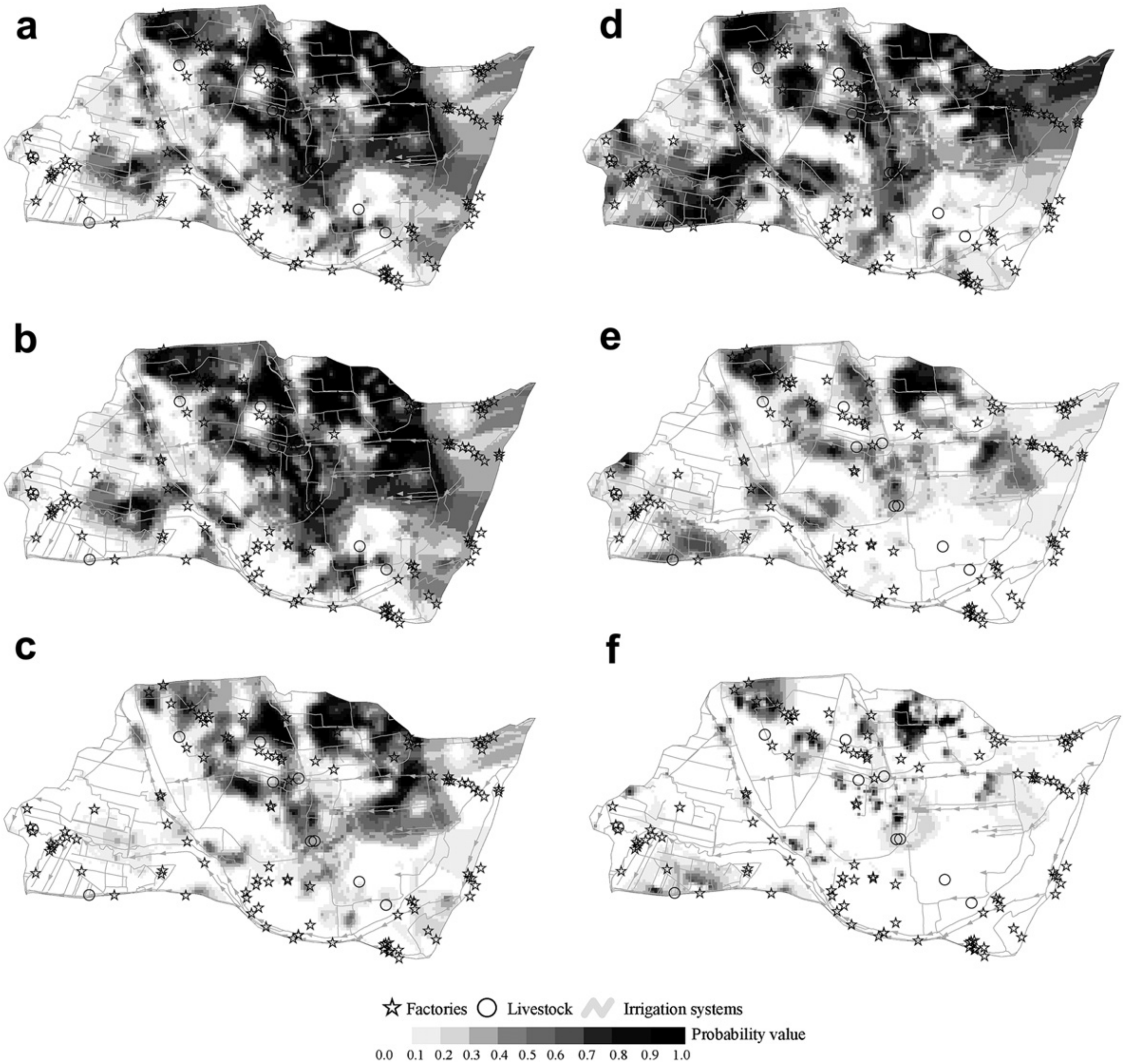

Fig. 5. Indicator kriged probability maps based on (a) regulatory threshold of Ni = 200 mg kg^{-1}, (b) FMDM cut-off value Ni = 193.2 mg kg^{-1}, (c) FMDM cut-off value Ni = 290.2 mg kg^{-1}, (d) FMDM cut-off value Zn = 353.5 mg kg^{-1}, (e) regulatory threshold of Zn = 600 mg kg^{-1}, (f) FMDM cut-off value Zn = 963.0 mg kg^{-1}. (FMDM: finite mixture distribution model).

Cr and Cu soil pollution, as well as moderate Ni and Zn pollution. According to the probability maps (Fig. 6c and 6d), the entire study area should be considered at high risk of pollution when Cr, Cu, Ni or Zn exceeds the regulatory thresholds or FMDM's moderate pollution cut-off values. Single-variable indicator kriging with regulatory thresholds and the FMDM cut-off values provides a valuable reference for efforts to determine the environmental action in order to mitigate heavy metal soil pollution. Notably, MVIK provides more conservative information and scenarios than single-variable indicator kriging does. Moreover, the probability maps of single-variable indicator kriging and MVIK with the FMDM cut-off values and the regulation thresholds of heavy metals in soil provide accurate and alternative information for identifying soil pollution sources, pathways, high probability polluted areas, as well as areas requiring remedial environmental action.

3.7. *Validation of soil contaminations*

Since the ROC curve is a standard means of evaluating the goodness of fit of a model (Manel et al., 2001; Swets, 1986) and the forecasting accuracy. Values above 0.7 are generally considered good while values exceeding 0.9 are considered as an excellent model forecasting or classification capability. Contamination is accurately forecasted when the ROC value is one. The contamination classification of IK and MVIK associated with heavy metal pollution thresholds in soils is assessed and validated using the ROC curve. The ROC validation results in Table 5 indicate that the ROC values for the indicator kriging with FMDM cut-off values range from 0.97 to 1. Moreover, IK and MVIK have excellent capabilities to classify contaminations (ROC area >0.94). MVIK with FMDM classifications of soil pollutions are highly reliable (ROC area >0.96) for both conditional probability exceeding cut-off values and exceeding any cut-off values, but with regulatory threshold of MVIK is 0.94 slightly less than 0.96. IK with FMDM threshold of Zn is 1.0, indicating that FMDM cut-off values of 963 mg kg^{-1} 100% can classify high amounts of Zn contamination in soils accurately. The ROC validation results further demonstrate that FMDM cut-off values with indicator and multiple indicator kriging can reliably delineate heavy metal contamination in soil, particular for areas with high heavy metal concentrations in soil.

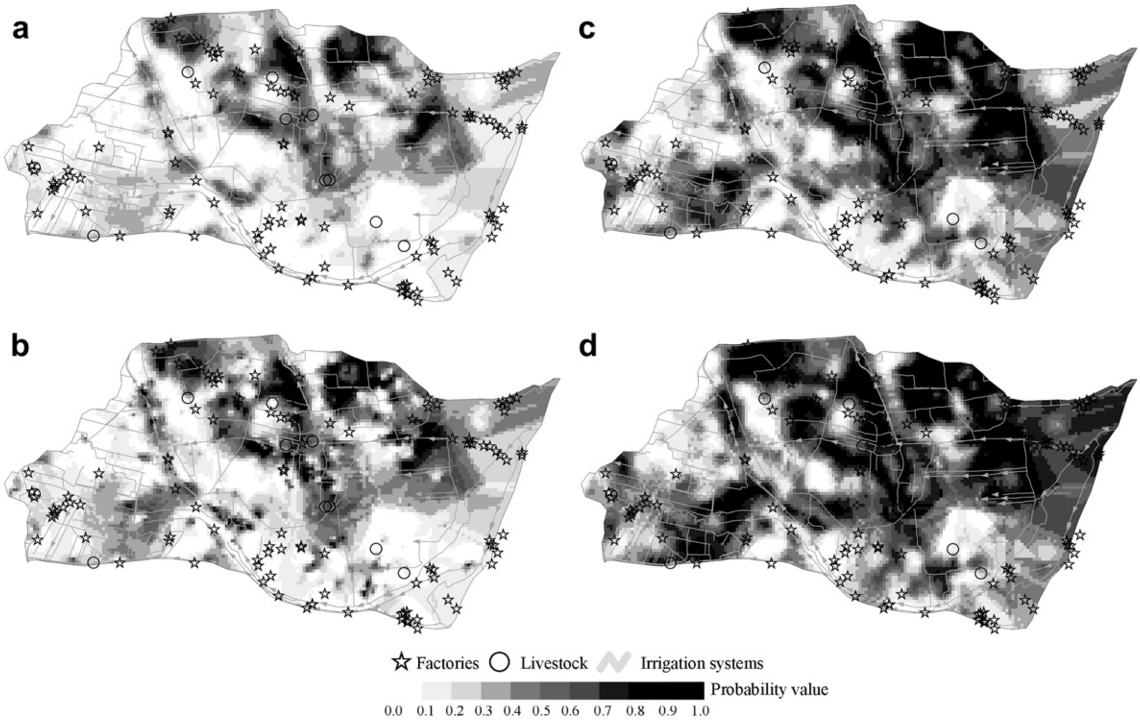

Fig. 6. MVIK (a) conditional probability map of Cr, Cu, Ni and Zn exceeding RT, (b) conditional probability map of Cr, Cu, Ni and Zn exceeding FMDM cut-off values, (c) probability maps for areas exceeding any RT, (d) probability maps for areas exceeding any FMDM cut-off values. (FMDM: finite mixture distribution model).

Table 5
ROC values of contamination validations.

	IK						MVIK	
	Cr	Cu	Ni		Zn		Conditional probability exceeding cut-off values	Exceeding any cut-off values
Cu-off values	0.982	0.985	0.988	0.969	0.971	1.000	0.986	0.961
RT	0.993	0.986	0.988		0.964		0.941	0.958

Cut-off values: cut-off values of finite mixture models.
RT: regulatory thresholds.
IK: indicator kriging.
MVIK: multiple-variables indicator kriging.

4. Conclusions

Evaluating the probable polluted spatial patterns, the background values and the anthropogenic origin *versus* the natural origin of heavy metals in soils is essential to assessing the impact of human activities and providing accurate information for remedial environmental actions. While proposing a novel approach that integrates FMDM, single-variable indicator kriging and MVIK, this study determines the cut-off values of heavy metal pollutants in soil, as well as identifies pollution sources, pathways and areas in the study region. Based on samples taken from mixtures of polluted and unpolluted area without understanding the full areal extent of the contamination, the FMDM cut-off values identify the characteristics of the statistical distributions, as well as the natural and anthropogenic background of Cr, Cu, Ni and Zn in soil in the study area. The values of the background concentrations of Cr, Cu, Ni and Zn can provide a valuable reference, along with additional samples, for determining the background of heavy metals in soil in the study area. The values also provide alternative contamination levels for moderate and high heavy metal pollution in the soil. Single-variable indicator kriging with the regulation thresholds and the FMDM cut-off vales provides accurate information and useful scenarios for identifying probable pollution sources, pathways and areas with specific heavy metal pollutants in the soil. MVIK with the FMDM cut-off values and the regulation thresholds provides more conservative information and scenarios than the single-variable indicator kriging for environmental planning and management does. The proposed approach can be used not only to confirm and redefine pollution thresholds, but also to identify areas where the soil is probably polluted by heavy metals in order to facilitate local environmental remediation efforts. We recommend that a future study assess the uncertainty of delineating soil contamination by undertaking spatial uncertainty analysis with FMDM.

References

Alatorre, L.C., Beguería, 2009. Identification of eroded areas using remote sensing in a badlands landscape on marls in the central Spanish Pyrenees. Catena 76, 182–190.
Albanese, S., Vivo, B.D., Lima, A., Cicchella, D., 2007. Geochemical background and baseline values of toxic elements in stream sediments of Campania region (Italy). Journal of Geochemical Exploration 93, 21–34.
Alloway, B.J., 1995. Heavy Metals in Soil, second ed. Blackie Academic & Professional, UK.
Brus, D.J., de Gruijter, J.J., Walvoort, D.J.J., de Vries, F., Bronswijk, J.J.B., Romkens, P.F.A.M., de Vries, W., 2002. Mapping the probability of exceeding critical thresholds for cadmium concentrations in soils in the Netherlands. Journal of Environmental Quality 31, 1875–1884.
Cattle, J.A., McBratney, A.B., Minasny, B., 2002. Kriging method evaluation for assessing the spatial distribution of urban soil lead contamination. Journal of Environmental Quality 31 (5), 1576–1588.
Chen, M., Ma, L.Q., Harris, W.G., 1999. Baseline concentrations of 15 trace elements in Florida surface soils. Journal of Environmental Quality 28, 1173–1181.
Chen, T., Liu, X., Li, X., Zhao, K., Zhang, J., Xu, J., Shi, J., Dahlgren, R.A., 2009. Heavy metal sources identification and sampling uncertainty analysis in a field-scale vegetable soil of Hangzhou, China. Environmental Pollution 157, 1003–1010.

Chen, Z.S., 2000. Relationship between heavy metal concentrations in soils of Taiwan and uptake by crops. Food and Fertilizer Technology Center – Technical Bulletin 149, 1–15.

Cressie, 1993. Statistics for Spatial Data. Wiley, New York.

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

Du, J., 2002. Combined algorithms for constrained estimation of finite mixture distributions with grouped data and conditional data. M.S. Thesis, McMaster University, Hamilton, Canada, 124 pp.

Dudka, 1993. Base-line concentrations of As, Co, Cr, Cu, Ga, Mn, Ni and Se in surface soils, Poland. Applied Geochemistry Suppl. 2, 23–28. Environmental Geochemistry.

El Sebai, T., Lagacherie, B., Soulas, G., Martin-Laurent, F., 2007. Spatial variability of isoproturon mineralizing activity within an agricultural field: geostatistical analysis of simple physicochemical and microbiological soil parameters. Environmental Pollution 145, 680–690.

Facchinelli, A., Sacchi, E., Mallen, L., 2001. Multivariate statistical and GIS based approach to identify heavy metal sources in soils. Environmental Pollution 114, 313–324.

Folkes, D.J., Kuehster, T.E., Litle, R.A., 2001. Contributions of pesticide use to urban background concentrations of arsenic in Denver, Colorado, U.S.A. Environmental Forensics 2, 127–139.

Huang, S.S., Liao, Q.L., Hua, M., Wu, X.M., Bi, K.S., Yan, C.Y., Chen, B., Zhang, X.Y., 2007. Survey of heavy metal pollution and assessment of agricultural soil in Yangzhong district, Jiangsu Province, China. Chemosphere 67, 2148–2155.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

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, 229–238.

Juang, K.W., Lee, D.Y., Ellsworth, T.R., 2001. Using rank-order geostatistics for spatial interpolation of highly skewed data in a heavy-metal contaminated site. Journal of Environmental Quality 30, 894–903.

Kiekens, L., 1995. In: Alloway, B.J. (Ed.), Zinc, Heavy Metals in Soils, second ed. Blackie Academic & Professional, UK, pp. 284–305.

Krebs, C.J., 1999. Ecological Methodology, second ed. Benjamin Cummings, Menlo Park, California.

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

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

Liu, C., Zhang, L., Davis, C.J., Solomon, D.S., Gove, J.H., 2002. A finite mixture model for characterizing the diameter distributions of mixed-species forest stands. Forest Science 48, 653–661.

MacDonald, P.D.M., Pitcher, T.J., 1979. Age groups from size-frequency data: a versatile and efficient method of analyzing distribution mixtures. Journal of the Fishery Research Board of Canada 36, 987–1001.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.

Manta, D.S., Angelone, M., Bellanca, A., Neri, R., Sprovieri, M., 2002. Heavy metals in urban soils: a case study from the city of Palermo (Sicily), Italy. Science of the Total Environment 300, 229–243.

Martín, J.A.R., Arias, M.L., Corbí, J.M.G., 2006. Heavy metals contents in agricultural topsoils in the Ebro basin (Spain). Application of the multivariate geostatistical methods to study spatial variations. Environmental Pollution 144, 1001–1012.

Martínez, J., Llamas, J., de Miguel, E., Rey, J., Hidalgo, M.C., 2007. Determination of the geochemical background in a metal mining site example of the mining district of Linares (South Spain). Journal of Geochemical Exploration 94, 19–29.

McGrath, S.P., 1995. Chromium and nickel. In: Alloway, B.J. (Ed.), Heavy Metals in Soils, second ed. Blackie Academic & Professional, UK, pp. 152–178.

McLachlan, G., Peel, D., 2000. Finite Mixture Models, first ed. Wiley, New York, USA.

Micó, C., Recatalá, L., Peris, M., Sánchez, J., 2006. Assessing heavy metal sources in agricultural soils of an European Mediterranean area by multivariate analysis. Chemosphere 65, 863–872.

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

Portier, K.M., 2001. Statistical issues in assessing anthropogenic background for arsenic. Environmental Forensics 2, 155–160.

Rodríguez, J.A., Nanos, N., Grau, J.M., Gil, L., López-Arias, M., 2008. Multiscale analysis of heavy metal contents in Spanish agricultural topsoils. Chemosphere 70, 1085–1096.

Schnabel, U., Tietje, O., Scholz, R.W., 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.

Swets, J.A., 1986. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions, first ed. Wiley, New York, USA.

Ungaro, F., Ragazzi, F., Cappellin, R., Giandon, P., 2008. Arsenic concentration in the soils of the Brenta Plain (Northern Italy): mapping the probability of exceeding contamination thresholds. Journal of Geochemical Exploration 96, 117–131.

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

Verstraete, S., Van Meirvenne, M., 2008. A multi-stage sampling strategy for the delineation of soil pollution in a contaminated. Environmental Pollution 154, 184–191.

Wang, H., 2002. Assessment and prediction of overall environmental quality of Zhuzhou City, Hunan Province, China. Journal of Environmental Management 66, 329–340.

Yang, S.Y., Chang, W.L., 2005. Use of finite mixture distribution theory to determine the criteria of cadmium concentrations in Taiwan farmland soils. Soil Science 170 (1), 55–62.

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.I., 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, 1423–1433.