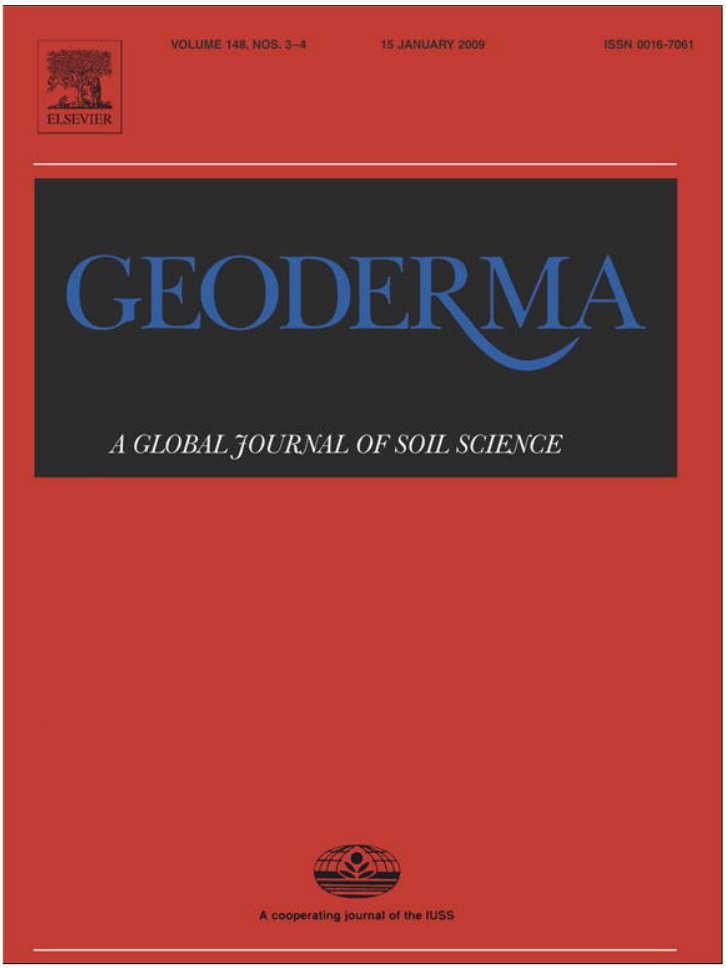

Geoderma 148 (2009) 413–420

Contents lists available at ScienceDirect

Geoderma

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

Combining marginal and spatial outliers identification to optimize the mapping of the regional geochemical baseline concentration of soil heavy metals

T. Meklit [a,*], M. Van Meirvenne [a], S. Verstraete [a], J. Bonroy [a], F. Tack [b]

[a] Gent University, Faculty of Bioscience Engineering, Department of Soil Management, Coupure 653, 9000, Gent, Belgium
[b] Gent University, Faculty of Bioscience Engineering, Department of Applied Analytical and Physical Chemistry, Coupure 653, 9000, Gent, Belgium

ARTICLE INFO

Article history:
Received 20 May 2008
Received in revised form 18 September 2008
Accepted 14 November 2008
Available online 3 December 2008

Keywords:
Geochemical baseline concentration
Marginal outliers
Spatial outliers
Fuzzy classification
Robust variogram
Chromium

ABSTRACT

The geochemical baseline concentration is used as a reference to determine the state of an area in relation to soil pollution. Various methods have been developed to determine this concentration based on filtering either the marginal or the spatial outliers. Marginal outlier identification (MOI) classifies data as belonging to the geochemical baseline or representing pollution using a globally defined single threshold value. As a result it neglects the local scale variability of the geochemical baseline level that arises from possible differences in parent material and the presence of multiple pollutants with variable degrees of influence. Hence it might lead to the identification of enrichments below the globally defined threshold but still larger than the local geochemical baseline level as belonging to the geochemical baseline. Spatial outlier identification (SOI) focuses on detecting unusual values in a local neighbourhood. As SOI is strongly dependent on data configuration, clusters of high values might wrongly be accepted as being geochemical baseline data that can inflate geochemical baseline level in pollution risk areas. The limitations of MOI and SOI can be severe when applied for a large scale study. To avoid these limitations and maximize the benefit of the two methods we proposed a combined methodology: integrated outliers identification (IOI) using fuzzy and robust means to determine the geochemical baseline measurements of Cr for Flanders, Belgium. Through the use of IOI it was possible to identify both scattered and clustered outliers resulting in determination of Cr geochemical baseline level that does not deny the local as well as the regional scale variability and display a higher degree of spatial structure as expected for the geochemical baseline data.

1. Introduction

The concentration of heavy metals in soils is highly variable due to a number of natural and anthropological reasons. The original level from the parent material can be altered at different scales by bio-geochemical process, anthropological activities and depositions of aerially transported particles. When the concentration exceeds a given reference value, the soil is considered to be contaminated. In soil quality assessment studies, to determine the state of a specific area with regard to soil pollution therefore a reference concentration, mostly called 'background concentration', is required.

However, the definition of a background concentration is not clear cut. Pfannkuch (1990) defined it as "the natural base load of an element", Porteous (1996) referred to it as "a concentration that would exist without a local polluting source", while the other definition in Reimann and Garrett (2005) puts it as "the concentration that can no longer be proven to originate from a polluting source". An important disagreement between these definitions is whether an enrichment that is caused by a diffused contamination should be considered as background concentration or not.

As an alternative, scientists were prompted to use 'geochemical baseline concentration' as a reference. This term was first introduced by the international geochemical mapping programs to describe the current variation in concentration of an element in the surficial environment (Salminen and Tarvianen, 1997). It integrates the geochemical background and diffuse contamination (Sierra et al., 2007). The geochemical baseline measurements do not claim to be the natural background because anthropological influences are accounted for next to the influence of biogeochemical processes (Kabata-Pendias and Pendias, 1984; Salminen and Gregorauskiene, 2000).

Since Europe has a long history of industrialization and agriculture we believe that the anthropological factor needs to be considered to characterize the variability in soil heavy metal geochemical concentrations of the region. This paper deals with the determination of the geochemical baseline concentration to be used as reference for soil pollution studies. Although the term 'baseline' can be sometimes understood as a single threshold value (Reimann and Garrett, 2005) that differentiates non-polluted from polluted soils for the whole study area, in this paper the geochemical baseline concentration refers to the concentration range that can vary at the regional scale but

* Corresponding author. Tel.: +32 9 264 60 42; fax: +32 9 264 62 47.
E-mail address: Meklit.tarikuchernet@ugent.be (T. Meklit).

0016-7061/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.geoderma.2008.11.014

describes a more homogeneous distribution at a local scale. The variability at a regional scale is a result of the large size of our study area where the anthropological factor contribution and the presence of multiple pollution sources with different degrees of influence can be importantly variable.

The following model was used to describe the nature of our data:

$$Z(\mathbf{x}) = G(\mathbf{x})(1-r(\mathbf{x})) + P(\mathbf{x})r(\mathbf{x}) \quad (1)$$

where $Z(\mathbf{x})$ is an observation of variable Z at location $\mathbf{x}$, $G(\mathbf{x})$ is a random function describing a continuous autocorrelated process of the geochemical baseline concentration, $P(\mathbf{x})$ is a concentration that is a result of contaminating process referred as pollution data and $r(\mathbf{x})$ is an indicator variable for the rate of contamination with a range of values between [0,1]. The two extreme cases are obtained when $r(\mathbf{x})$ takes the value of one or zero. The first instance is a result of heavy pollution where observations reflect no more the geochemical baseline concentration of the area but rather pollution measurements. In the absence of artificial deposition of the material in the environment $r(\mathbf{x})$ takes a value of zero and observations reflect the geochemical concentration. In reality however the change in concentration between the two extremes is gradual which is then explained by $r(\mathbf{x})$ values between 0 and 1.

Since the geochemical baseline concentration is the result of the combination of the native metal content in the soil parent material and the deposition of diffuse contamination, at a local scale it describes the concentration range that is commonly found (Tack et al., 1997). Pollution data originating from point sources like industries, in contrast, are characterized by a few numbers of observations often with extremely high concentrations. These limited number of high concentration measurements form contaminated patches with a finite spatial extent around the source (Lark, 2002).

As a result the geochemical baseline and the pollution data are different in both their statistical distribution and spatial behaviour. The pollution data behave as marginal outlier with respect to the overall data distribution (Rawlins et al., 2005). But due to the unusual value they have within their local neighbourhood the pollution data can also be considered as spatial outliers (Lark, 2002).

Based on these differences methods have been developed to determine the geochemical baseline concentration. The available methods are built on the principle of filtering out either the marginal or the spatial outliers.

A measurement is considered as a spatial outlier if it has a value that is eccentric in its local neighbourhood. Such a local outliers can be a result of faulty measurements or it can also be the actual value of an unrepresentative sample such as from relocated soil. Marginal outliers are observations of highly polluted areas; these measurements can be actual values but do not represent the geochemical baseline concentration of the study area. By the term “outlier” we therefore are not directly questioning the genuineness of the value but whether it is coherent with the local as well as the regional data distribution.

Methods that are based on marginal outlier identification (MOI) usually classify data into geochemical baseline and pollution data using a globally defined threshold derived from the overall statistical distribution. This is mostly conducted using a graphical representation, such as a probability plot, where the inflexion on the graph is considered as a separation point of the two populations (Fleischhauer and Korte, 1990; Tobías et al., 1997; Tack et al., 2005; Sierra et al., 2007).

Spatial outlier identification (SOI), on the other hand, focuses on detecting records with unusual values in their local neighbourhood. Variograms are used to model the spatial autocorrelation and with a cross-validation procedure of ordinary kriging an estimated value is generated for every measurement. Bárdossy and Kundzewich (1990) and Laslett and McBratney (1990) then used the standardized estimation error as a criterion to identify the spatial outliers.

At a regional scale, which is the situation addressed in this paper, on top of the natural variability of the element, the fluctuation of man-induced contributions and presence of multiple pollution sources with variable degrees of influence resulted in variable geochemical baseline levels. Under such circumstances the marginal outlier identification based on the global distribution of the data does not guarantee a unique definition of the geochemical baseline concentration at a local scale. Traceable sources leading to enrichments below the dividing threshold value but beyond the local geochemical baseline levels will still be included for the establishment of the geochemical baseline concentrations.

Data used for a regional scale study often have important differences in sampling density. It is a common practice to collect large number of samples in pollution risk zones that result in the formation of clusters of high values. This is a challenge for the implementation of the SOI procedure to determine the geochemical baseline concentration correctly. Since the identification of the spatial continuity depends on the sampling density there is a risk of accepting clusters of extremely large values as the geochemical baseline measurements. As a result the geochemical baseline concentration in pollution prone areas can be exaggerated.

The determination of the regional geochemical baseline concentration has to be however efficient regardless of scale, local versus regional, and sampling density. To achieve this objective we propose to integrate MOI and SOI. Our approach is called integrated outliers identification (IOI). Through filtering both the marginal and the spatial outliers the risk of accepting clustered high values is avoided while the definition of the geochemical baseline level at a local scale remains valid. Our approach is demonstrated using a soil chromium (Cr) database from Flanders, Belgium. The main objective of this paper is to evaluate different approaches for outlier identification and removal prior to further processing (like the delineation of polluted areas).

2. Materials and methods

2.1. Study area and soil samples

The study area covers the entire region of Flanders (13,522 km^2) being the northern part of Belgium. The soil is dominantly developed in eolian or marine sediments of Holocene and Pleistocene age (Van Meirvenne and Van Cleemput, 2005). The northern part of the region is dominated by acid, humus rich sandy soils. Finer wind-blown sediments were deposited in the southern parts, resulting in loamy and silty textures.

The Cr data used in this study were obtained from the Public Waste Agency of Flanders (OVAM), the regulatory institute responsible for waste management and soil remediation in Flanders. OVAM requires a soil evaluation during the transfer of land or when an area is suspected of being polluted. Due to the obligation to analyse Cr in every soil sample, even if no pollution by heavy metals is expected, this database contains both Cr concentrations below and above the geochemical baseline. In Belgium, emissions of Cr from ferrous industries were documented by Thiessen et al. (1988).

The Cr data were collected between 1996 and 2005. During this time OVAM has been giving assignments for different laboratories to collect and analyze soil samples for heavy metals. The samples were analyzed according to the standard procedure of OVAM and the required procedure for total Cr analysis involves microwave destruction of 0.5 g of the air dry fine-earth fraction (<2 mm) of soil with 6 ml 37% HCl, 2 ml 65% HNO_3 and 2 ml 40% HF (OVAM, 1992; method CMA/2/II/A.3). In the digest Cr was analyzed by ICP-AES (OVAM, 1992; method CMA/2/I/B.1). Although all laboratories followed this standard procedure some differences in the detection limit of used measuring equipments was observed. The majority of the measurement equipment has a detection limit around 5 mg kg^{-1} of Cr while some have a limit as low as 0.02 mg kg^{-1}.

Every measurement was located by its geographical coordinates, upper and lower sampling depth and sampling date. Due to the presence of samplings with identical coordinates and due to variations in sampling depth, the database had to be screened carefully. In the case where repeated data were taken at the same location, the most recent was retained. For reasons related to arable soil use, in this study we targeted the top 50 cm of the soil profile. After screening, 14,458 observations were available to represent the Cr concentration in the top 50 cm of the soil of Flanders (Fig. 1). The entire region was covered, but more intense sampling occurred around large urban and industrial areas (like Antwerp, Gent and Kortrijk). Because of the uneven spatial density, the need for declustering the data to obtain an unbiased population distribution was checked using a cell declustering algorithm (Deutsch and Journel, 1998).

2.2. Marginal outlier identification

Geochemical baseline data are largely influenced by the geological parent material and are commonly observed concentration levels at a local scale within the study area. Pollution data on the contrary often contain a limited number of extremely large concentrations (Hirotaka and Goovaerts, 2000). As a result these two datasets show differences in statistical distribution, the latter behaving as marginal outliers in the overall statistical distribution. Some authors claim that the geochemical baseline measurements follow a lognormal distribution (Fleischhauer and Korte, 1990; Salminen and Gregorauskiene, 2000) and accordingly they used a normal probability plot using the logarithmically transformed data to identify the two datasets. The probability plot is chosen since on such a plot a normal distribution is presented as a straight line while the skewed tail creates a deviation from linearity and forms a bend on the curve. So the identification of the marginal outliers can be done by removing observations subsequently, starting with the largest concentration, until the remaining part of the distribution is linear (i.e. normal). As a measure of normality the coefficient of skewness, which is zero for normally distributed data, has been used. This procedure has been applied in several studies, including Fleischhauer and Korte (1990), Tobías et al. (1997) and Tack et al. (2005). However, the assumption of a lognormal distribution for the geochemical baseline data is criticized (Reimann and Filzmoser, 2000; Sierra et al., 2007). The use of the point of least skew as a dividing threshold is also considered to be biased.

Mostly there are many and often superimposed factors responsible for the distribution of heavy metals in soil. Thus, the evolution from the geochemical baseline to the pollution data could be expected to be gradual, which is a phenomenon explained in the model when $r(\mathbf{x})=(0,1)$ (Eq. (1)). Therefore it is unlikely to assume all observations to fall into one of the two groups. With available techniques however there is no room to address observations that are neither geochemical baseline nor pollution data. For the statistical tools to be used it is therefore necessary not to be dependent on the assumption of lognormal distribution for the geochemical baseline data and to be able to address observations that do not fit in to one of the two datasets. To serve this purposes our choice of the tool was the Fuzzy *k*-means with extragrades where the membership information can be used to assess the degree of belongingness of every observation for a specific class and outliers that are outside the different classes can also be identified (McBratney and de Gruijter, 1992).

The fuzzy *k*-means classification method has been used successfully to classify soil properties (McBratney and Moore, 1985; Vitharana et al., 2008). This algorithm assigns a partial membership value, μ_{ic}, for every measurement, i, to be a member of an individual class, c, by minimizing iteratively the objective function, $J(\mathbf{M},\mathbf{C})$ (Odeh and McBratney, 1992). For data that contain n objects ($i=1,\ldots,n$) with p-attributes ($v=1,\ldots,p$) which must be classified into k-classes ($c=1,\ldots,k$), the minimization of the objective function can be expressed as:

$$J(\mathbf{M},\mathbf{C}) = \sum_{i=1}^{n}\sum_{c=1}^{k} \mu_{ic}^{\phi} d_{ic}^{2}(\mathbf{x}_i,\mathbf{c}_c) \qquad (2)$$

where $\mathbf{M}=\mu_{ic}$ is the matrix of membership values and $\mathbf{C}$ is the centroid of class c for variable v, $\mathbf{x}_i$ is the vector representing object i, $\mathbf{c}_c$ is the vector representing the centroid of class c, is the square distance between $\mathbf{x}_i$ and $\mathbf{c}_c$ according to an Euclidian, Mahalonobis or diagonal distance metric. ϕ is the fuzziness exponent which controls the degree of fuzziness of the classification. It can take a value between 1 and ∞. A

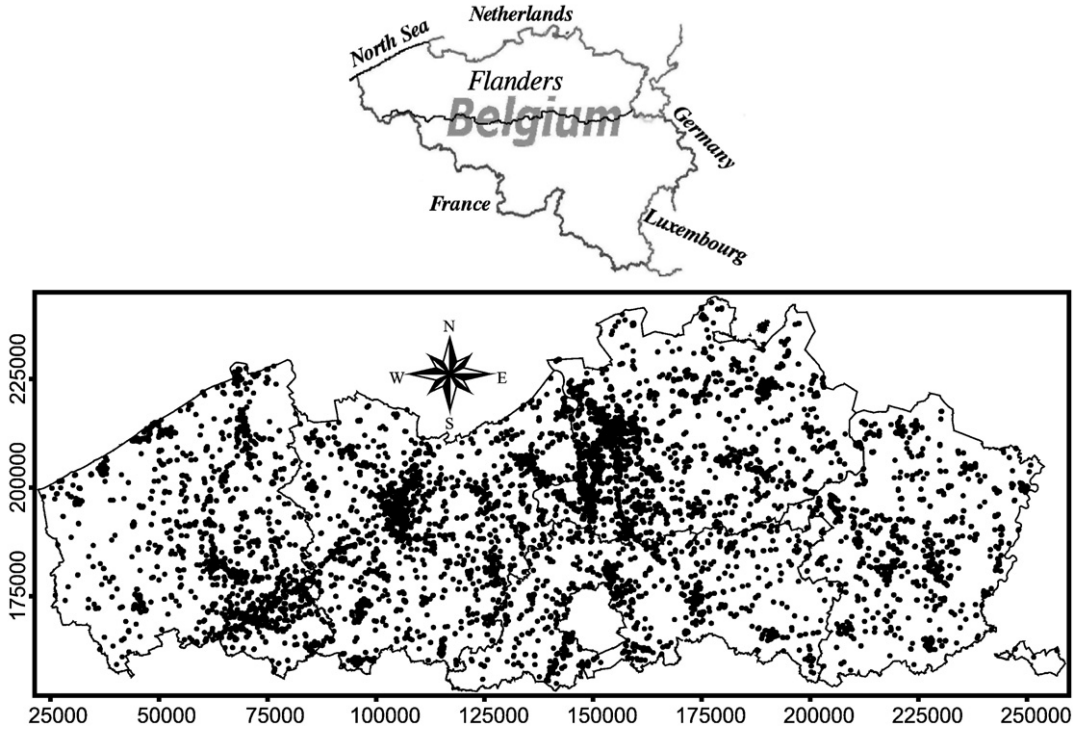

Fig. 1. Belgium with identification of Flanders (top) and the 14458 available Cr sampling locations (bottom), coordinates are in m according to the Belgian Lambert72 projection.

crisp classification results from $\phi=1$, which results in membership values being either 0 or 1. On the other hand, a large ϕ value results in clusters with an almost identical members.

By making the memberships directly depend upon the distances to the class centroids the objective function J was modified to accounts for extragrades (McBratney and de Gruijter, 1992) as follows:

$$J(\mathbf{M},\mathbf{C}) = \beta \sum_{i=1}^{n}\sum_{c=1}^{k} \mu_{ic}^{\phi} d_{ic}^{2}(\mathbf{x}_i,\mathbf{c}_c) + (1+\beta)\sum_{i=1}^{n} \mu^{*\phi} \sum_{c=1}^{k} d_{ic}^{-2} \quad (3)$$

where μ^* is the membership to a fuzzy class of extragrades and β is a parameter that determines the mean value of μ^*.

Fuzzy k-means with extragrades was chosen for its capacity to identify and assign the extragrades in a special class. Extragrades are observations that are outside the conventional classes (McBratney and de Gruijter, 1992), which are then observations that belong to neither the geochemical baseline nor the pollution data.

The iterative minimization of $J(\mathbf{M},\mathbf{C})$ demands a decision on the number of classes, k; the fuzziness exponent, ϕ; a distance metric for d_{ij}^2; and a stopping criteria, e. For the determination of optimum number of classes and the fuzziness exponent, combinations of Fuzziness Performance Index (FPI), Modified Partition Entropy (MPE) and the negative derivative of $J(\mathbf{M},\mathbf{C})$, with respect to ϕ are used (Triantafilis et al., 2003). For the optimal number of classes however it is also a common practice to use the user's knowledge of the data (Gorsevski et al., 2003). Since our objective is to classify our data in the way that they explain the major controlling factors of soil heavy metal distribution, i.e., either the instances of the geochemical baseline or the pollution data we defined the value of k to be 2. Since the number of classes are defined, the determination of the value of ϕ was done following the scheme proposed by McBratney and Moore (1985), by calculating $J(\mathbf{M},\mathbf{C})$ values for a series of ϕ and plotting the curve of the negative derivative of $J(\mathbf{M},\mathbf{C})$, i.e., $-[(\delta J/\delta\phi)k^{0.5}]$ versus ϕ where the optimal ϕ value was obtained at the maximum of this curve. The function of $\delta J/\delta\phi$ was defined by Bezdek (1980) as:

$$\frac{\delta J}{\delta \phi} = \sum_{i=1}^{n}\sum_{c=1}^{k} \mu_{ic}^{\phi} log(\mu_{ic}) d_{ic}^{2}. \quad (4)$$

For d_{ic}^2 the diagonal distance matrix was used and the stopping criterion was set at 0.0001 to end the iteration when the difference between consecutive membership matrixes dropped below this value.

After determining the optimum fuzziness exponent the Cr data were processed with fuzzy k-means with extragrades where every observation obtained membership values for each of the three different classes: the geochemical baseline data class, the pollution data class and the extragrade class. An observation was finally assigned to a class for which its membership was largest.

2.3. Spatial outlier identification

The identification of spatial outliers (i.e. a large value compared to the local observation surrounding it) can be done using the standardized estimation error, $\varepsilon_s(\mathbf{x}_0)$, as criterion (Bárdossy and Kundzewich, 1990; Laslett and McBratney, 1990):

$$\varepsilon_s(\mathbf{x}_0) = \frac{z^*(\mathbf{x}_0) - z(\mathbf{x}_0)}{\sigma(\mathbf{x}_0)} \quad (5)$$

where $z^*(\mathbf{x}_0)$ is the estimated value at $\mathbf{x}_0$ and $\sigma^2(\mathbf{x}_0)$ is the kriging variance at $\mathbf{x}_0$. This method requires modelling the spatial autocorrelation by the variogram and the kriging estimation at every sampling location in turn, assuming not to know the measurement (known as the cross-validation procedure). For a spatial outlier the difference between the estimated value and the true value can be expected to be strongly negative, due to the large difference with neighbouring observations. Additionally, in such situations it is also likely that the variance is underestimated by $\sigma^2(\mathbf{x}_0)$ (Lark, 2002). Consequently, a large standardized error can be expected at the localisation of spatial outliers. As evaluation criterion, Bárdossy and Kundzewich (1990) and Laslett and McBratney (1990) labelled an observation as a spatial outlier if $\varepsilon_s(\mathbf{x}_0)$ is smaller than −1.96.

For the modelling of the variogram, typically the Matheron variogram (Matheron, 1962) is used. This classical variogram considers the squared difference between two measured values:

$$\gamma_M(\mathbf{h}) = \frac{1}{2N(\mathbf{h})}\sum_{\alpha=1}^{N(\mathbf{h})} \{z(\mathbf{x}_\alpha) - z(\mathbf{x}_\alpha+\mathbf{h})\}^2 \quad (6)$$

where $\gamma_M(\mathbf{h})$ is the Matheron variogram, $z(\mathbf{x}_\alpha)$ and $z(\mathbf{x}_\alpha+\mathbf{h})$ are observations separated by a distance vector $\mathbf{h}$ and $N(\mathbf{h})$ is the number of pairs $(z(\mathbf{x}_\alpha), z(\mathbf{x}_\alpha+\mathbf{h}))$.

In the presence of outliers in the data, due to the possible large differences in values, the Matheron variogram estimator is limited for proper modelling of the spatial correlation; it can become unstable (Cressie, 1993). Therefore, to work with a large dataset that contains spatial outliers, Lark (2000) suggested the use of the more robust Dowd variogram (Dowd, 1984). The Dowd estimator takes the median of the absolute pair differences of $z(\mathbf{x}_\alpha)$ and $z(\mathbf{x}_\alpha+\mathbf{h})$ as a basis for the variogram estimator:

$$2\gamma_D(\mathbf{h}) = 2.198^*\{\text{median}(|y_\alpha(\mathbf{h})|)\}^2 \quad (7)$$

where $\gamma_D(\mathbf{h})$ is the Dowd variogram and $y_\alpha(\mathbf{h}) = \{z(\mathbf{x}_\alpha) - z(\mathbf{x}_\alpha+\mathbf{h})\}$. The value 2.198 is a correction factor that scales the median absolute deviation to correspond with the standard deviation of normally distributed data.

Using the robust variogram is an appropriate way to find spatial outliers. However, the robust estimators are less efficient than the Matheron's in the absence of outliers. Thus, the decision of which estimator to use needs to be first made after checking the presence of outliers in the data. For this purpose Lark (2000) introduced a statistics $\theta(\mathbf{x})$ that can be used to test the median of the standardized squared kriging errors.

$$\theta(\mathbf{x}) = \frac{\{z^*(\mathbf{x}_0) - z(\mathbf{x}_0)\}^2}{\sigma^2\mathbf{x}_0} \quad (8)$$

where $z^*(\mathbf{x}_0) - z(\mathbf{x}_0)$ is the difference between estimated and measured values using a cross-validation and $\sigma^2\mathbf{x}_0$ is the kriging variance.

Lark (2002) showed that if we krige an intrinsic data using a proper variogram, the $\theta(\mathbf{x})$ will have a χ^2 distribution with one degree of freedom and hence the median has a value of 0.455. If there are no spatial outliers in the data, the median of $\theta(\mathbf{x})$ using the Matheron variogram estimator will not be significantly different from 0.455 and if that is not the case the best variogram estimator to be selected will be the one with median $\theta(\mathbf{x})$ closest to 0.455.

2.4. Integrated outlier identification

The integrated outlier identification (IOI) which we propose combines the MOI and SOI methods to identify both the marginal and spatial outliers in this sequence. Although the Dowd variogram is quite robust to the influence of outliers in the dataset, local extreme values still have their impact. Therefore, we propose IOI where first the marginal outliers were identified with MOI. Next, the SOI procedure was applied on the remaining data. The $\varepsilon_s(\mathbf{x}_0)$ statistic was finally used to exclude the spatial outliers.

2.5. Modelling the variogram

For all variograms the model fitting was done separately based on an iterative minimization of the sum of the squared differences

between the experimental values and the theoretical model, taking into account the number of data pairs and the lag distance (Pardo-Igúzquiza, 1999). A double exponential model (Eq. (9)) was found to fit to the data best:

$$\gamma(\mathbf{h}) = C_0 + C_1 * \left(1 - exp\left(-\frac{3\mathbf{h}}{a_1}\right)\right) + C_2 * \left(1 - exp\left(-\frac{3\mathbf{h}}{a_2}\right)\right) \quad \forall \quad 0 < \mathbf{h} \qquad (9)$$
$$\gamma(0) = 0$$

where C_0 is the nugget variance representing unstructured or short-distance variability, the sum of C_0, C_1 and C_2 represents the total variability, called the sill, and a_1 and a_2 are the two range parameters. The nugget-to-sill ratio, NSR=$C_0/(C_0+C_1+C_2)$, was used as an indication of the strength of autocorrelation of variable Z.

3. Results and discussion

3.1. General statistics of Cr data

The result of declustering showed that the mean reduces with increasing cell dimension indicating the presence of preferential sampling in areas with high values of Cr concentration. Considering the declustering weight assigned for every observations the regional mean Cr concentration was found to be 36.4 mg kg^{-1} with a median of 29.2 m kg^{-1} (Table 1). The data showed a characteristic skewed distribution with a wide range, from 0 to 29300 mg kg^{-1}, where 95% of the readings having a value of less than 80 mg kg^{-1}.

3.2. Marginal outlier identification

The Cr data were classified into outliers, the geochemical baseline and the pollution data by a fuzzy k-means with extragrades algorithm using the FuzMe software (Minasny and McBratney, 2006). To identify the optimal fuzzy exponent value, the objective function was calculated for ϕ= 1.1,…, 2.8 in steps of 0.1. The curve of the derivative of the objective function versus ϕ (Fig. 2) indicated that for k=2 the optimal ϕ value was 2.5.

Finally, 11 155 observations with a range of 5–45 mg kg^{-1} Cr concentration were assigned to the class which was considered as the MOI geochemical baseline. According to the geochemical atlas of Europe by Salminen et al. (2005) for Flanders the maximum of the top soil geochemical Cr concentration varies between 28 and 44 mg kg^{-1}. The Vlaamse Gemeenschap (1996) gives the background concentration of Cr for a standard soil of Flanders, defined as a soil containing 10% of clay and 2% of organic matter, to be 37 mg kg^{-1}. The result we obtained from the fuzzy k-means classification is quite similar with the atlas information. The difference between the average value of the regional background concentration given by Vlaamse Gemeenschap and the fuzzy k-means result is also acceptable considering a prevailing soil variation in the region.

The summary statistics of these 11155 data are shown in Table 1. As a result of excluding the marginal outliers the regional mean, the quartiles and the median of the data reduced markedly. Since observations with high Cr measurements are excluded from the data obviously, the variance reduced dramatically.

Table 1
The summary statistics of the whole data set and those resulting after the different steps of removing outliers

	Whole data	MOI	SOI	IOI
Number of data	14458	11155	13590	10697
Mean	36.4	25.1	32.7	24.7
Variance	16256.3	108.2	852.6	106.0
Maximum	29300	45.0	1603.6	45.0
Upper quartile	42.9	33.0	40.8	33.0
Median	29.2	25.0	28.2	24
Lower quartile	19.0	17.0	18.9	16.4
Minimum	0.02	5.0	0.02	5.0

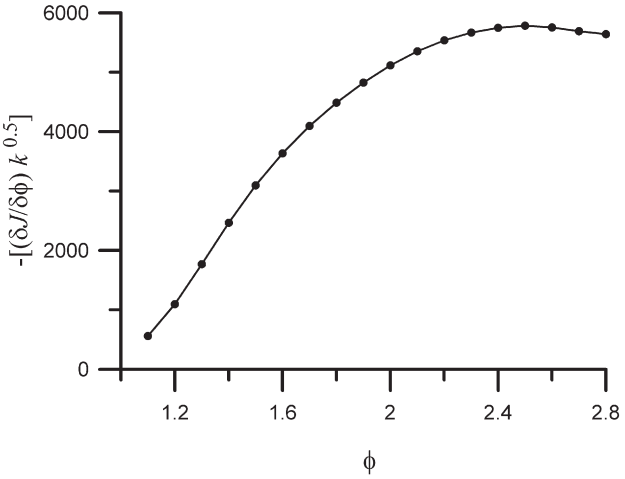

Fig. 2. Determination of the optimal fuzziness exponent ϕ for k=2.

3.3. Spatial outlier identification

In order to choose the proper variogram that describes the spatial autocorrelation best, the Matheron as well as the Dowd variograms were modeled from the 14447 Cr measurements. The median $\theta(\mathbf{x})$ of the Matheron variogram was found to be 0.19, which is significantly different from the ideal 0.455 value. The Dowd estimator on the other hand had a median $\theta(\mathbf{x})$ of 0.40 hence this variogram was used to identify spatial outliers.

Subsequently, all the Cr measurements were used to calculate the Dowd variogram with Eq. (6) and a double exponential model was fit to it. The result is given by Fig. 3. With the parameters of the Dowd variogram, ordinary kriging was used to estimate Cr values at every observation location, neglecting every measurement in turn (cross-validation). Next the standardized estimation error, $\varepsilon_s(\mathbf{x}_0)$, was computed to identify spatial outliers. In total 857 data were identified as spatial outliers. After removing them the remaining 13590 observations constituted the SOI geochemical baseline dataset. With this procedure the change in summary statistics of the data in general is weaker as compared to the result of MOI (Table 1). This result indicates that applying SOI does not focus on a removal of high or low values but on values that are unusual in their neighbourhood.

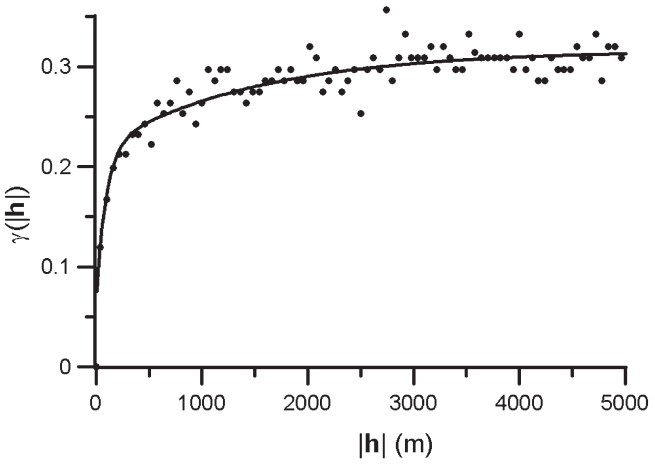

Fig. 3. The Dowd variogram calculated with the whole Cr dataset, modelled with double exponential structure with C_0=0.07, C_1=0.15, a_1=279 m, C_2=0.1 and a_2=4500 m.

In Flanders, the maximum critical sanitation threshold of Cr requiring remediation measures to be implemented in industrial areas is 800 mg kg^{-1}. However, after excluding the spatial outliers, Cr concentrations above this threshold, and up to 1603 mg kg^{-1}, were retained among the geochemical baseline measurements. These large values were located in industrial site where there are clusters of extreme high values and so they were not recognized as spatial outliers. This clearly is unacceptable and illustrates the shortcoming of this approach.

When the spatial distribution of the observations retained as the geochemical baseline data by MOI and SOI are compared, some interesting differences appear. Observations which were rejected as marginal outliers by MOI but accepted as the geochemical baseline data by the SOI method occurred in clusters of high values. As a result of the dependency of the spatial outlier identification method on the sampling configuration, even clustered high values were accepted as geochemical baseline data. Thus for data with an important difference in sampling density the SOI method is inadequate. It is however important to note the effectiveness of the SOI technique for excluding scattered high valued data which deviate strongly from the values in their neighbourhood.

Observations which were accepted after excluding the marginal outliers, but which were rejected as spatial outliers, were found scattered over the region containing relatively elevated Cr concentrations as compared to observations in their neighbourhood. As a consequence, local enrichments below the globally defined threshold value were included in the geochemical baseline dataset. However, MOI was efficient in excluding outliers independent of the sampling density.

3.4. Integrated outlier identification

After removing the marginal outliers, the presence of outliers in the remaining 11155 observation was again checked by calculating the median $\theta(\mathbf{x})$ both for Matheron and Dowd estimators, which resulted in 0.23 and 0.43, respectively. As a result the Dowd estimator was used to model the spatial autocorrelation of these observations and a double exponential model was fit to it (Fig. 4). Ordinary kriging was used to estimate the Cr values at every observation location, neglecting every measurement in turn and the standardized estimation error was computed. In this way 458 data were identified as spatial outliers, so with IOI method 10697 observations were accepted as the geochemical baseline data. The summary statistics of these data is also given in Table 1. In general there is a slight difference between the results of IOI and MOI where the mean, median, and lower quartile values of the later are somewhat smaller. On the contrary the difference between IOI and SOI is quite large.

To examine the degree of autocorrelation, the Matheron variograms (Eq. (5)) of the data identified as the geochemical baseline by the three methods (MOI, SOI and IOI) were calculated. Table 2 provides the parameters of the double exponential model which was used to model these variograms. After removing the marginal outliers, the data resulted from MOI has the NSR of 0.44. The result of SOI shows a strong decrease in the NSR, 0.21. A stronger reduction in the proportion of the unstructured variability of the data was the result of the identification and removal of outliers from the dataset in terms local information. As compared to the SOI, the integrated approach, IOI, although does not further reduce the NSR, it resulted in a clear increase of the long-range (a_2) parameter. So after excluding the spatial and the marginal outliers, the remaining data displayed a stronger spatial structure as could be expected for geochemical baseline data since they are expected to behave according to gradual patterns related to geological processes and diffused pollution within the study area.

Table 2
The parameters of the Matheron variograms using the data accepted as the geochemical baseline by the different methods

	MOI	SOI	IOI
C_0	0.09	0.07	0.04
C_1	0.09	0.23	0.13
a_1 (m)	400	550	470
C_2	0.03	0.04	0.03
a_2 (m)	5000	7000	8210
NSR	0.44	0.21	0.20

In order to asses the risk of not removing the outliers on the estimation; the three datasets were analyzed further using their respective Matheron variograms parameter. Since our objective was to compare between estimated values that can be obtained from the three techniques a straightforward interpolation method, ordinary kriging, was chosen. The resulting maps (not shown since mapping was not our main objective) show differences at a local scale mainly. To compare the result, 45 mg kg^{-1} Cr was used as a threshold to determine the area with estimations above this value. With MOI over 280 km^2 area received estimates above this threshold, with a maximum estimated concentration of 56 mg kg^{-1}. With SOI this area reduced to 223 km^2, with a maximum of 87 mg kg^{-1}, and with IOI the area reduced further to 178 km^2 with a maximum estimated value of 52 mg kg^{-1}. It is clear that the different methods had different impacts on the spatial estimations based on the data selected by each method.

3.5. Estimation of the local geochemical baseline concentration

The Matheron variogram of the geochemical baseline data obtained by the IOI approach (Fig. 5) is used to generate the Cr geochemical baseline map (Fig. 6) with Ordinary kriging procedure. A general trend

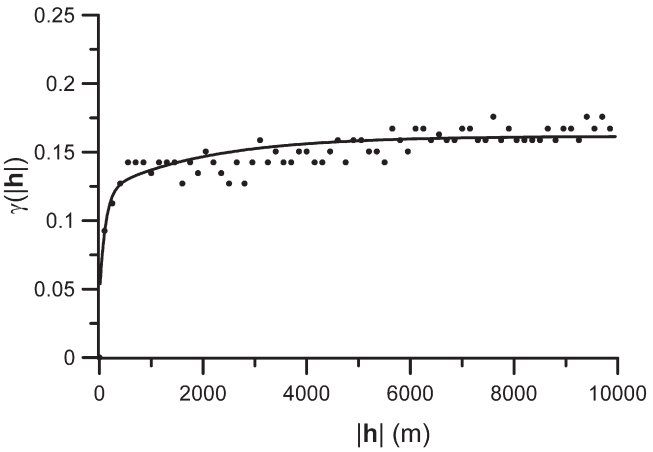

Fig. 4. The Dowd variogram of the 11155 data which were retained after removal of the marginal outliers and used to identify further the spatial outliers with IOI method; modeled with double exponential structure, C_0=0.04, C_1=0.09, a_1=339 m, C_2=0.04 and a_2=6010 m.

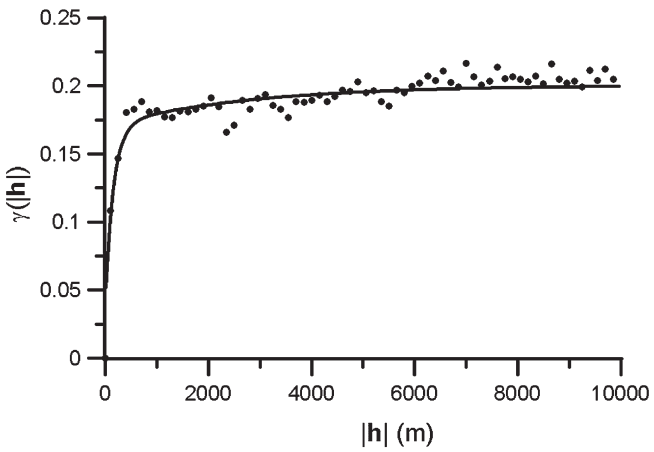

Fig. 5. The Matheron variogram of data defined as Cr geochemical baseline using IOI, modeled with double exponential structure, C_0=0.04, C_1=0.13, a_1=470 m, C_2=0.03 and a_2=8210 m.

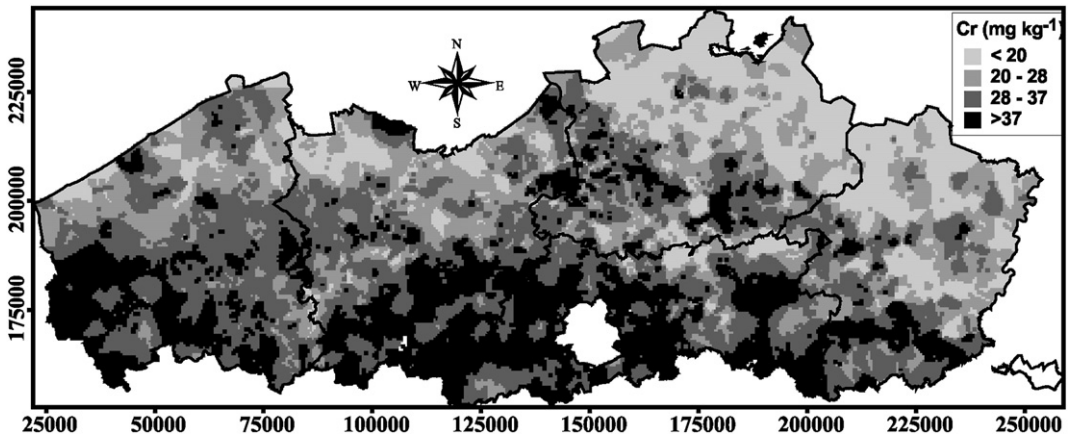

Fig. 6. Cr geochemical baseline distribution in Flanders as obtained from IOI, coordinates are in m according to the Belgian Lambert72 projection.

of increasing Cr concentration can be observed from north to south. Most of the northern part has a Cr geochemical baseline concentration of less than 20 mg kg^{-1}. But in the southern part geochemical baseline concentrations can reach up to 43 mg Cr kg^{-1}. This pattern reflects the general change in soil texture, i.e. from sandy in the north to the silty in the south. So there are indications that this variation is due to soil genesis processes. Locally however, like in the strongly industrialized area around Antwerp, this general pattern is disturbed. So it seems to be evident that the geochemical baseline is a result of both general natural processes occurring on a regional scale and, at a more local scale, human activities.

4. Conclusions

For the determination of the geochemical baseline concentration the use of fuzzy *k*-means with extragrades classification was found to be efficient for identification of the marginal outliers. With this procedure most observations with extreme high values were removed. However, because this approach does not take information about the neighbourhood of observations into account, it was incapable to define uniquely the geochemical baseline level at a local scale. As a result on the contrary of expectations for the geochemical baseline measurements the resulted data from MOI procedure displayed weaker spatial autocorrelation. SOI on the other hand, was very useful in improving the spatial autocorrelation within the geochemical baseline data as it excludes outliers based on local information. But since SOI depends on the sampling configuration of the data its ability in identifying outliers fails in clustered sampled areas which unfortunately coincide mostly with polluted sites. As a result SOI involves a risk of inflating the geochemical baseline concentration in pollution risk zones. By combining the two approaches our proposed approach, IOI, becomes capable of avoiding the problem of overestimating the local geochemical baseline level and maximizing the structured variability within the geochemical baseline data.

As a conclusion we recommend the IOI approach to identify the geochemical baseline of heavy metals on a regional scale.

Acknowledgment

We wish to thank the Public Waste Agency of Flanders (OVAM) for providing the data for this research.

References

Bárdossy, A., Kundzewich, Z.W., 1990. Geostatistical methods for the detection of spatial outliers in groundwater quality spatial fields. Journal of Hydrology 115, 343–359.
Bezdek, J.C., 1980. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.
Cressie, N., 1993. Statistics for Spatial Data. Wiley, New York. Revised edition.
Deutsch, C.V., Journel, A.G., 1998. GSLIB Geostatistical Software Library and User's Guide. Oxford University Press, New York.
Dowd, P.A., 1984. The variogram and kriging: robust and resistant estimators. In: Verly, G., David, M., Journel, A.G., Marechal, A. (Eds.), Geostatistics for Natural Resource Characterization, vol. 1. Reidel, Dordrecht 91–106.
Fleischhauer, H.L., Korte, N., 1990. Formulation of cleanup standards for trace elements with probability plots. Environmental management 14 (1), 95–105.
Gorsevski, P.V., Gessler, P.E., Jankowski, P., 2003. Integrating a fuzzy *k*-means classification and a Bayesian approach for spatial prediction of landslide hazard. Journal of geographical systems 5, 223–251.
Hirotaka, S., Goovaerts, P., 2000. Geostatistical interpolation of positively skewed and censored data in a dioxin-contaminated site. Environmental science and technology 34, 4228–4235.
Kabata-Pendias, A., Pendias, H., 1984. Trace Elements in Soils and Plants. CRC Press, Boca Raton, Florida.
Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51, 137–157.
Lark, R.M., 2002. Modeling complex soil properties as contaminated regionalized variables. Geoderma 106, 173–190.
Laslett, G.M., McBratney, A.B., 1990. Further comparison of spatial methods for predicting soil-pH. Soil science society of America Journal 54, 1553–1558.
Matheron, G., 1962. Traité de Géostatistics Appliqué, Tome 1. Memoires du Bureau de Recherches Geologiques et Minières, Paris.
McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy K-means with extragrades. Journal of Soil Science 43, 159–175.
McBratney, A.B., Moore, A.W., 1985. Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology 35, 165–185.
Minasny, B., McBratney, A.B., 2006. *FuzME* version 3. Australian Centre for Precision Agriculture. The University of Sidney NSW.
Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1992. Soil pattern recognition with fuzzy *k*-means: application to classification and soil–landform interrelationship. Soil Science Society of American Journal 56, 505–516.
OVAM, 1992. Compendium voor Monsterneming en Analyse ter uitvoering van het afvalstoffendecreet en het bodemsaneringsdecreet. Openbare Afvalstoffenmaatschappij voor het Vlaamse Gewest, Mechelen.
Pardo-Igúzquiza, E., 1999. VARFIT: a fortran-77 program for fitting variogram models by weighted least squares. Computer & Geosciences 25, 251–261.
Pfannkuch, H.O., 1990. Elsevier's Dictionary of Environmental Hydrology. Elsevier, Amsterdam.
Porteous, A., 1996. Dictionary of Environmental Science and Technology, 2nd edn. J. Wiley, Chichester.
Rawlins, B.G., Lark, R.M., O'Donnell, K.E., Tye, A.M., Lister, T.R., 2005. The assessment of point and diffused metal pollution of soils from an urban geochemical survey of Sheffield, England. Soil use and management 21, 353–362.
Reimann, C., Filzmoser, P., 2000. Normal and lognormal distribution in geochemistry: death of myth. consequences for the statistical treatment of geochemical and environmental data. Environmental geology 39 (9), 1001–1014.
Reimann, C., Garrett, R.G., 2005. Geochemical background — concept and reality. Science of the total Environment 305, 12–27.
Salminen, R., Gregorauskiene, V., 2000. Considerations regarding the definition of a geochemical baseline of elements in the surficial materials in areas differing in basic geology. Applied Geochmistry 15, 647–653.
Salminen, R., Tarvianen, T., 1997. The problem of defining geochemical baselines: a case study of selected elements and geological materials in Finland. Journal of Geochemical Exploration 60, 91–98.
Salminen, R., Batista, M.J., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamic, J., Heitzmann, P., Lima, A., Jordan, G., Klaver, G., Klein, P., Lis, J., Locutura, J., Marsina, K., Mazreku, A., O'Connor, P.J., Olsson, S.Å., Ottesen, R.-T., Petersell, V., Plant, J.A., Reeder, S., Salpeteur, I., Sandström, H., Siewers, U., Steenfelt, A. & Tarvainen, T., 2005. Geochemical Atlas of Europe. Part 1: Background Information, Methodology and Maps. Geological Survey of Finland, Espoo [http://www.gtk.fi/publ/foregsatlas, accessed on 22/6/2007].

Sierra, M., Martínez, F.J., Aguilar, J., 2007. Baseline for trace elements and evaluation of environmental risk in soils of Almería (SE Spain). Geoderma 139, 209–219.
Tack, F.M.G., Verloo, M.G., Vanmechelen, L., Van Ranst, E., 1997. Geochemical baseline concentration levels of trace elements as a function of clay and organic carbon contents in soils in Flanders (Belgium). The Science of the Total Environment 201, 113–123.
Tack, F.M.G., Vanhaesebroeck, T., Verloo, M.G., Van Rompey, K., Van Ranst, E., 2005. Mercury baseline levels in Flemish soils (Belgium). Environmental pollution 134, 173–179.
Thiessen, L.M., Hallez, S., Lenelle, Y., Verduyn, G., 1988. Evaluation of the total Cr-levels in the ambient air in Belgium. The science of the total environment 71, 519–526.
Tobías, F.J., Bech, J., Sánchez, A.P., 1997. Statistical approach to discriminate background and anthropogenic input of trace elements in soils of Catalonia, Spain. Water Air and Soil Pollution 100, 63–78.
Triantafilis, J., Odeh, I.O.A., Minasny, B., McBratney, A.B., 2003. Elucidation of physiographic and hydrological features of the lower Namoi valley using fuzzy k-means classification of EM34 data. Environmental modelling and software 18, 667–680.
Van Meirvenne, M., Van Cleemput, I., 2005. Pedometrical techniques for soil texture mapping at a regional scale. In: Grunwald, S. (Ed.), Environmental Soil-Landscape Modeling. Geographical Information Technologies and Pedometrics. CRC Press, Taylor & Francis Group. Boca Raton, FL, USA. 323–341.
Vitharana, U.W.A., Van Meirvenne, M., Simpson, D., Cockx, L., De Baerdemaeker, J., 2008. Key soil and topographic properties to delineate potential management classes for precision agriculture in the European loess area. Geoderma 143, 206–215.
Vlaamse Gemeenschap, 1996. Besluit van de Vlaamse regering houdende vaststelling van het Vlaams reglement betreffende de bodemsanering. Belgisch Staatsblad of 27.03.1996, Brussels.