USE OF SOFT DATA IN A GIS TO IMPROVE ESTIMATION OF THE VOLUME OF CONTAMINATED SOIL

L. A. M. HENDRIKS[1], H. LEUMMENS[1], A. STEIN[1] and P. J. DE BRUIJN[2]
[1] *Department of Soil Science and Geology, Agricultural University, P.O. Box 37, 6700 AA Wageningen, The Netherlands;* [2] *DHV Environment and Infrastructure, P.O. Box 1076, 3800 BB Amersfoort, The Netherlands*
E-mail: alfred.stein@bodlan.beng.wau.nl

(Received 8 November 1995; accepted 25 September 1996)

Abstract. In the practice of soil remediation, organoleptic observations such as the smell or the colour of contaminated soil play an important role when determining well-defined volumes of contaminated soil. A GIS is then used to combine quantitative measurements with such soft data. In this study general procedures concerning how to deal with this type of observations are presented. The procedures were applied to a former gas works site, which was contaminated with cyanide and polycyclic aromatic hydrocarbons in the Netherlands. The volume of contaminated soil was determined. Use of soft data reduced the uncertainty in the volume of contaminated soil with 4 to 16%.

Key words: cyanide, geostatistics, GIS, PAH, soft data

1. Introduction

Geographical Information Systems (GIS) are suitable to a variety of environmental problems. They may be used to delineate sub-areas that have to be isolated and/or remediated, to inspect visually the quality of the sampling pattern and to obtain insight into the effects of different sanitation techniques. Their use is most advantageous when many different sources of information need to be combined. Until now emphasis has been placed on combined historical information, soil maps, satellite data and point observations (Stein, 1994). In the majority of these studies, the point observations were considered to be hard data: measurements were made in well-equipped laboratories. Geostatistical interpolation procedures (kriging) were then of a great value for extending this information to locations where measurements were not made. In particular, stratification based on the soil map or on historical information as well as the use of co-variables such as elevation provided useful procedures to include as much relevant information as possible for interpolation (Staritsky *et al.*, 1992; Leenaers *et al.*, 1989). Interactive GIS may be used to improve the processing of data by accounting for knowledge that is otherwise unaccessible, for example in intelligent Geographical Information Systems (Burrough, 1992; Stein *et al.*, 1995).

In many environmental studies organoleptic observations related to texture, colour and smell of soil samples are collected by experienced environmental workers to select appropriate sampling locations and hence minimize the costs of laboratory measurements. On the basis of these observations they may be able to

Water, Air, and Soil Pollution **101:** 217–234, 1998.

delineate parts of the area that are contaminated. In short: organoleptic observations can influence decisions for sanitation purposes substantially. In this study we identify organoleptic observations as soft data: they are often qualitative and are based on subjective judgement made on small volumes of soil which might have considerable consequences for environmental management. Until recently little has been known of how to deal with soft data in a GIS environment, for which points, lines and polygons are common types of objects. In mining and petroleum, however, use of soft data is already common (Zhu and Journel, 1993). This study focuses on the use of soft data in environmental studies. An additional problem encountered concerns the interpretation of these data from a statistical point of view and, in particular, how such data can be combined with hard data. In this study we used soft data with the aim of improving geostatistical procedures. The study is illustrated with an analysis of data from a gas work site in the eastern Netherlands.

2. Soft Data Related to Hard Data in a GIS

A common division of spatial objects in a GIS is the distinction between points, lines and polygons. Here it is assumed throughout that these objects have fixed attribute values (Molenaar, 1991): points are denoted by z_i, lines by l_j and polygons by p_k. In environmental soil studies it is a common practice to survey an area with augers from which soil samples are taken and analyzed. The measured on soil samples may be stored as attributes in a database (Rikkers *et al*., 1994) and as such are considered to be known with certainty. In this study attention is focused on soft data: these data are often qualitative and are subjectively observed. To use optimally soft data for data processing, a relation with hard data need to be established. To model the softness for points, it is assumed here that both the variable of interest, such as the concentration of a pollutant, and the soft data may be expressed by a random function $Z(x)$, depending on the location vector x in a 1-, 2- or 3-dimensional space. The actual observation (hard or soft) is a realization of the random function. Measurement errors are likely to be larger for soft data. Such differences in the patterns of spatial continuity could be accounted for by using cokriging (Dowd, 1993). In principle the concept of softness extends to lines and polygons as well (Burrough, 1989) but these will not be treated in the current study.

Three different forms of soft point data are distinguished (Journal, 1986; Zhu and Journal, 1993):

1. Prior information: prior information, such as historical information concerning the presence of building, may be helpful to delineate homogeneous sub-areas; roads and rivers are lines, l_j, and the delineations of the sites for old buildings are polygons, p_k.
2. Indicator data: at a certain location an observation is known to exceed a critical value without needing to take a measurement. For example, in case of mineral oils, the dark colour of the soil and strong smell may indicate sharply that the

particular spot is polluted: a specified threshold value, z_a is exceeded: $Z(x) \geq z_a$, without specifying *how much* the value z_a is exceeded. Interval data indicating that at a certain location an observation falls between some specified, i.e. $z_a \leq Z(x) \leq z_b$ can be treated in a similar fashion. Slight blue colouring of a soil polluted with cyanide may indicate a small degree of pollution, which is likely to be less severe than that indicated by dark blue colouring. Sometimes a particular distribution, such as a log-normal distribution, is assumed to model this type of uncertainty. In most situations, however, the form and the size of the distribution cannot be inferred from a single observation.

3. Indirect data: based on the presence or absence of physical objects, inference is made concerning the most likely value of the variable to be studied. Typical examples in environmental studies include remnants of buildings and bricks to which relatively high values of contaminants are associated.

In this paper a 3D-volume of polluted soil is described using values from chemical analyses and organoleptic observations. Measurement are located in R^3 using the vector of coordinates (x_1, x_2, x_3). The object is stored in a single database in a GIS.

3. Study Area

The study area of 5 ha on a former gas works site is located in the city of Enschede in the eastern Netherlands (Figure 1). The site was used to obtain gas from coal, starting in the late nineteenth century until the middle of the nineteen sixties. As a result of the various industrial processes and handling of by-products and waste substances such as tar, aromatic organic compounds, ammonical liquors and cyanide, by human activities and soil forming processes, the distribution of contaminants is very heterogeneous. Coke residues are found on the gas work sites mixed with soil. Prussian Blue (iron sulphides and ferric ferrocyanide) used in the purification of the coal gas before it was stored in gas tanks is also found on the site. It was decided to use a GIS to store the available information and to visualize the volume of contaminated soil, including its uncertainties.

Within the immediate perimeter of the premises at the gas works site so-called key-areas were defined and encircled within a radius of 25 m around each of the described buildings. On the basis of historical information, four relevant spatial objects were identified as key areas:

- Gas was stored in large *gas tanks*, in which condensates accumulated on the (unpaved) floors, in particular Naphtalene and Cyanide.
- Tar and ammonia were stored in *tar and ammonia tanks*; in addition tar was also stored in a *coal-tar tank*: both tar and ammonia became available in large quantities during gas production.
- Spent oxide was dumped in a *spent oxide building*. When gaining gas from coal, hydrated iron oxide, used to extract sulphate and cyanide from the gas,

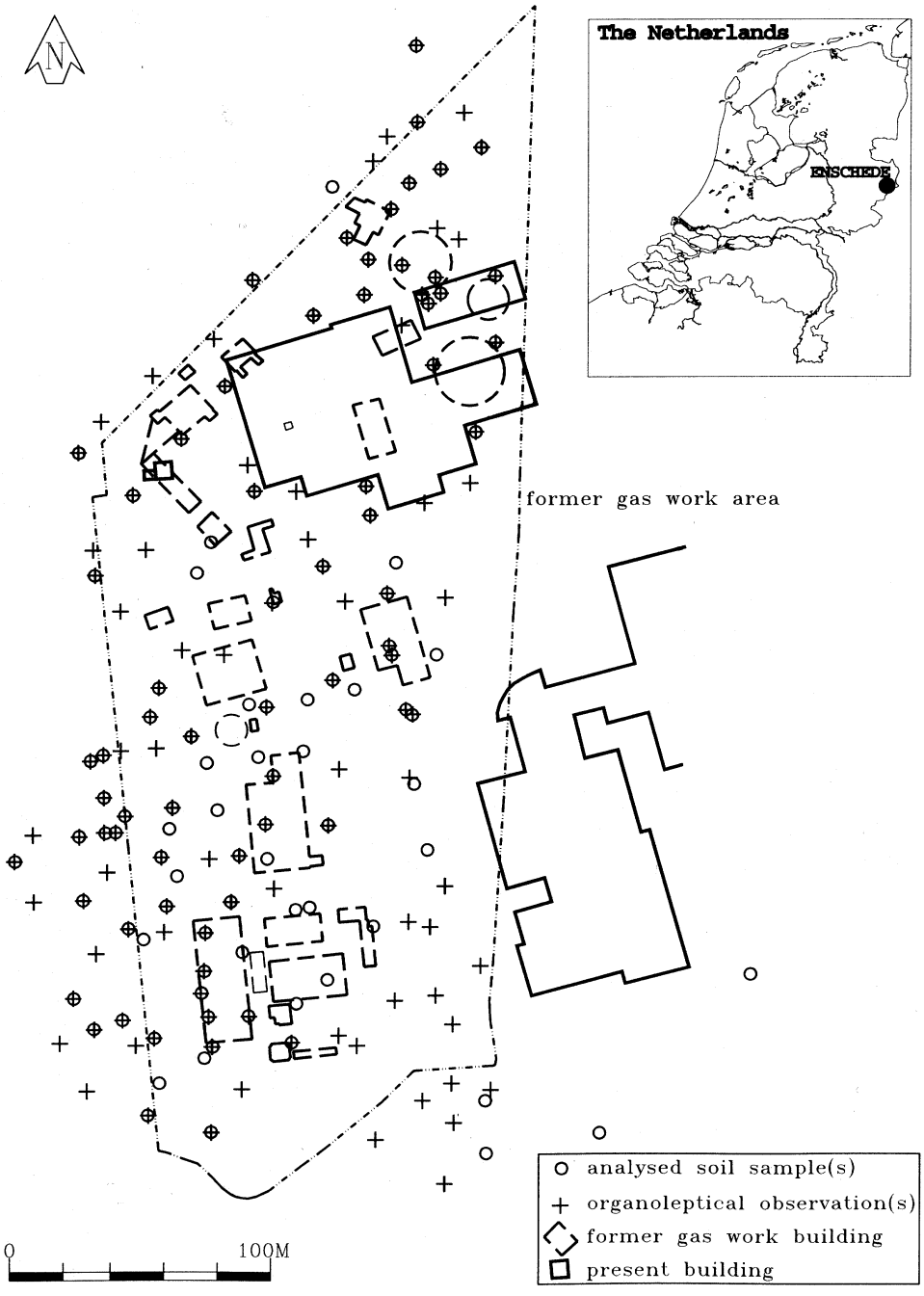

Figure 1. Study area located in the eastern Netherlands. Marked are the observation locations and previous buildings.

is periodically regenerated by exposing it to air or by blowing air through it. After the sulphur content reached a level of 45–50%, the iron oxide became useless. It was either stored, or it was sold as a weed killer. Due to cyanide complexes, dumped (spent) oxide is easily identified by its characteristic blue colour and smell.

– The *outside area* is the area beyond the four other key areas.

This form of statification of the site helps to disinguish the possibly strongly polluted sub-areas from the possibly unpolluted areas on the basis of historical information. Only few quantitative analyses were made within the first three key areas: these areas were known to be strongly contaminated. The outside area was also little sampled, because of initial lack of interest in the supposedly non-contaminated area. There is, however, some contamination in the outside area due to (i) intensive anthropogenic movement of waste and contaminated materials during gas works operation, and the demolishing and levelling of the site and (ii) downward and subsequent lateral migration in the groundwater.

For the three-dimensional volume of soil three layers were defined on the basis of hydrological conditions of the area. The anthropogenic top-layer (0–2 m below the soil surface) is permanently above groundwater. The second layer (2–4 m below the soil surface) is alternately above groundwater and below the highest groundwater level. The third layer, consisting of sand to 7 m below the soil surface, followed by a loam layer of 1–2 m thickness and again sand to 20 m below the soil surface, overlaying massive tertiary clay, is permanently below groundwater.

The two major contaminants, i.e. tar and spent oxide, have a large content of the polycyclic aromatic hydrocarbons (ΣPAH) and cyanide (*CN*), respectively. The ΣPAH consist of a list of 10 different *PAHs* (VROM, 1990). For this study attention was focused on pollution by ΣPAH and *CN* in the soil. It is common practice to define for each contaminant so-called target and intervention-levels: a concentration below the target-level indicates a clean observation, whereas an observation above the intervention-level indicates severe pollution requiring immediate action. Sometimes, also an intermedia value is used. The target, intermediate and intervention-levels for *CN* and ΣPAH for soil are 1, 10, 100 and 1, 20, 200 mg kg^{-1} dry matter, respectively.

The soil was sampled during successive soil surveys by means of borings. A soil sample and an organoleptic (soft) observation always characterize a vertical section of a 0.1 m diameter in the soil with a starting and a final depth. The locations of all observation points are shown in Figure 1. A total of 156 soil samples were analyzed for total *CN* and 91 soil samples for ΣPAH. At each location the soil core was described by partitioning the boring into separate depth intervals on the basis of organoleptic judgement. The number of depth intervals could be different for each boring, ranging from 1 and 2 (most commonly encountered) to 5. Clearly polluted depth intervals were not sampled, neither were clearly unpolluted depth intervals. From depth intervals where the degree of pollution was doubtful, a soil

sample was taken to determine type and degree of pollution in a laboratory. Each observation therefore relates to an organoleptically homogeneous depth interval.

In this study a general approach was formulated to derive depth data from organoleptic observations (Figure 2). Consider first the relatively simple situation of a single polluted layer. If at a certain location the measured ΣPAH or CN concentration is above the target-level, the depth to the upper surface U is equal to the starting depth of the sampling interval and the depth to the lower surface L to the final depth of the sampling interval. If the concentration is below the critical value, both U and L are set equal to the centre of the interval.

When two intervals are sampled three situations can be distinguished:

(i) both intervals contain contaminated soil: the depth to the first upper surface (U_1) was set to the starting depth of the first interval, the depth to the first lower surface (L_1) to the final depth of the first interval, the depth to the second upper surface (U_2) to the starting depth of the second interval and the depth to the second lower surface (L_2) to the final depth of the second interval;

(ii) one of the two intervals contains uncontaminated soil: either U_1 and L_1 or U_2 and L_2 are set equal to the centre of this interval, depending on which of the two intervals contains uncontaminated soil;

(iii) both intervals contain uncontaminated soil: both U_1 and L_1 as well as U_2 and L_2 are set equal to the centre of the first and the second sampling interval, respectively.

When a single interval was encountered in the neighbourhood of a double sampling interval, both U_1 and U_2 are set equal to U, and L_1 and L_2 to L. Extension of this procedure to more than two contaminated layers is straightforward, but was of little relevance in the current case study since only a few locations contained more than two sampling intervals. Note that sampled interval can be determined as polluted or not on the basis of either actual measurements, or organoleptic observations.

All soil survay observations that have been made since 1981 were recorded and stored in a database. Two types of organoleptic observations were recorded: one was measured very thoroughly, being based on the colour, smell and texture of the soil and a general description of the degree of contamination. The second form was measured roughly, being based mainly on smell: samples from which a strong smell was registered (at a safe distance) were likely to be more polluted than samples without smell. In addition, external visual presence of tar, tar odour, blue coloured soil and the presence of coal or coke residues were recorded, which may indicate a pollution with ΣPAH. CN contaminations are evident where the soil is blueish and has a particular almonry smell.

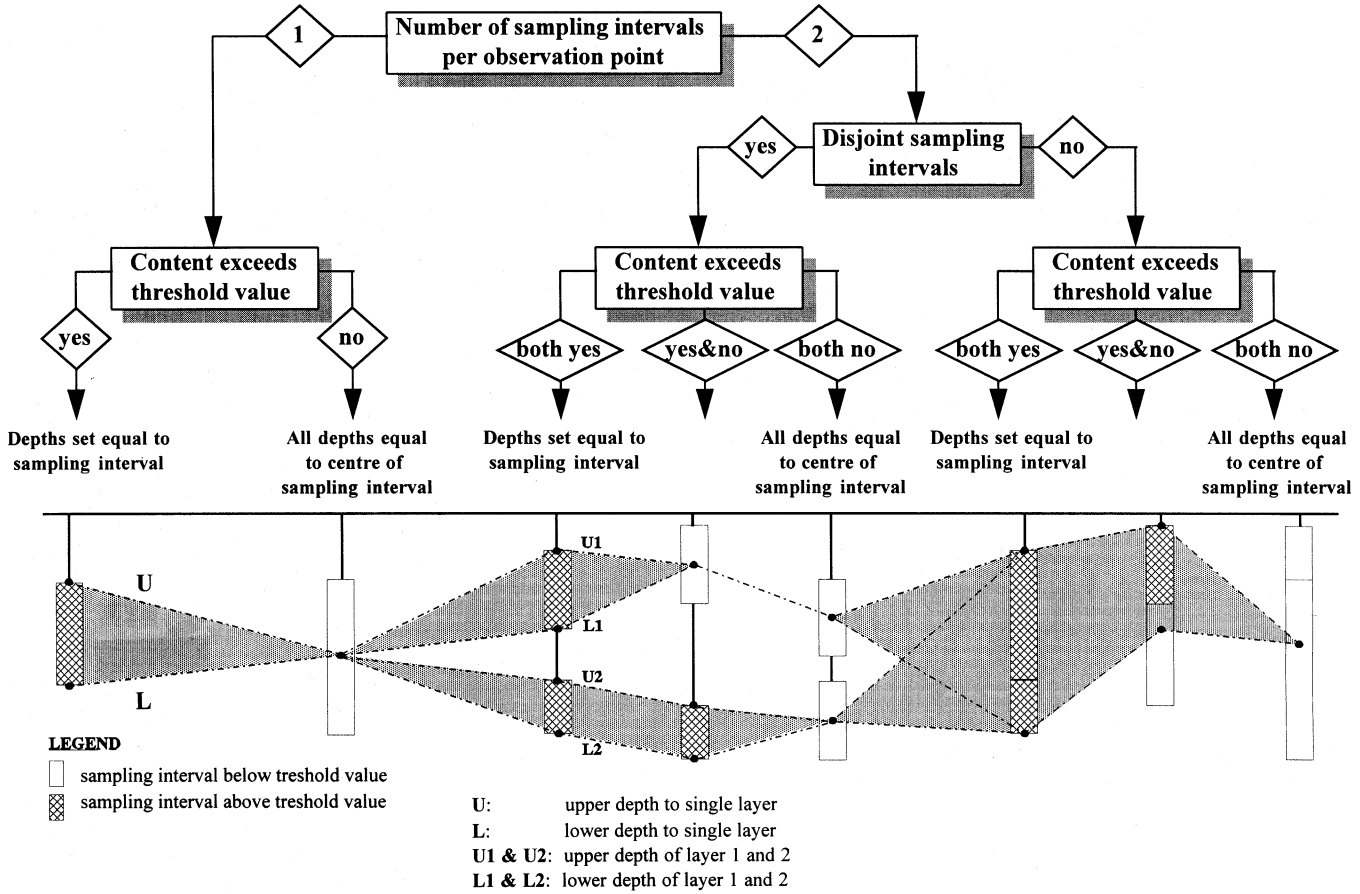

Figure 2. Definition of the uppe and lower depths to one or two polluted layers, using organoleptic observations within a GIS.

4. GIS and Geostatistics

An important option in GIS is to delineate three-dimensional objects from point observations. Geostatistical procedures are commonly used to predict the concentrations of a pollutant at a particular location, yielding also the associated uncertainties in form of the variance of the prediction error for these objects. The predicted values are stored in a GIS-database. This means that observations on several Z variables, Z_1, ..., Z_k have three coordinates (x_1,x_2,x_3) in a database (Raper and Kelk, 1991). Since a particular aim was to visualize the volume of contaminated soil including its prediction error, the database has to be accessed for displaying in a 3D representation, maybe after some necessary spatial transformations, for example to show the volume of soil where the predictions exceed the target-level. The GIS-database also has to be accessible in such a way as to be able to make cross-sections of the study area to give an optimal view of possibly contaminated locations.

To determine the spatial extent of polluted soil, we applied geostatistical techniques. Geostatistics is based upon regionalized variable theory (Matheron, 1965; Cressie, 1991). Measurements are viewed as realizations of random variables, which are tied to their measurement locations in a 1-, 2- or 3-dimensional space. The variables U, U_1, U_2, L, L_1 and L_2 for both *CN* and Σ*PAH* are considered as regionalized variables, and will be denoted by $U(x)$, $U_1(x)$, $U_2(x)$, $L(x)$, $L_1(x)$ and $L_2(x)$ to stress the relation with the location vector x. In most situations observations are spatially dependent, which means that observations that are close to each other in the ground are more alike than distant observations. This dependence is described by the variogram, which measures the degree of dissimilarity of the regionalized variable between places a distance h apart. The variogram for $U(x)$, $\gamma_u(h)$, is defined as

$$\gamma_u(h) = \frac{1}{2} E\left[(U(x) - U(x+h))^2\right]$$

where $U(x)$ and $U(x+h)$ denote the regionalized variable at two locations x and $x+h$, a distance h apart. The symbol *E[.]* represents mathematical expectation over the area of interest. For all the other variables a similar expression holds. In practice the variogram is computed as half the average squared difference between components of observed pairs:

$$\hat{\gamma}_u(h) = \frac{1}{2M(h)} \sum_{i=1}^{M(h)} (u(x_i) - u(x_i + h))^2$$

where $u(x_i)$ and $u(x_i+h)$ denotes a pair of measurements on the upper pollution surface, separated by a vector h, the total of such pairs being equal to $M(h)$. In many studies the variogram depends only on the length of the distance $|h|$ and not on the direction (isotropic case). A distinction will be made between the calculated

values $\hat{\gamma}_u(h)$, called the estimated semivariance, and the variogram, being its graph as a function of the separation distance $|h|$. In order to estimate reliably $\gamma(h)$, pairs of data are grouped into distance classes: pairs with approximately the same separation distance are used to estimate the semivariance for that particular distance, so that $M(h)$ is at lest 30 (Webster and Oliver, 1990). The length of a distance class is termed the lag.

Often the estimated semivariance for small value of h is small because the observations are fairly similar for this separating distance, whereas for a larger value of h larger differences are obtained. A mathematical model is usually fitted to the experimental variogram in order to duduce semivariance values for any possible distance required by interpolation procedures. A common variogram model is the exponential one, defined as:

$$g(h) = \begin{cases} C_0 + A^*(1 - e^{-\frac{|h|}{b}}) & \text{for} |h| > 0 \\ 0 & \text{for} |h| = 0 \end{cases}$$

where the parameters C_0 (the *nugget* variance), A (the *still* variance) and b (the *range*) may be estimated using a weighted non-linear regression procedure (Webster and Oliver, 1990), with weights equal to $M(h)$ for each distance. Other models could be chosen as well.

Kriging is a linear unbiased prediction of a value at an unvisited location by assigning weights λ_i to the observations $u(x_i)$, i = l,...,n:

$$\hat{u}(x_0) = \sum_{i=1}^{n} \lambda_i u(x_i)$$

with minimized prediction error variance (Cressie, 1991).

For both *CN* and Σ*PAH* variograms were computed for depths to the upper and lower boundaries of layer 1 and 2. A lag length of 10 m was used to have sufficient pairs of data points for estimation of semivariances. The depths to the layers were predicted using ordinary point kriging, with a neighbourhood of 8 observations. From the predicted values of upper and lower surfaces the total volume of contaminated soil was determined by multiplying the depths by the surfaces of the grid cells for both *CN* and Σ*PAH*. This is equal to the sum of the volume of the first polluted layer and the volume of the second contaminated layer, from which the intersection is subtracted. The total volume of contaminated soil is the soil that is either contaminated by *CN* or by Σ*PAH*. The uncertainty of the depth to each layer was determined by subtracting the kriging standard deviation from the depth to the upper boundaries and by adding it to the depth to the lower boundaries. This gives an approximate 68% bound for each depth. For each layer the uncertainty equals the sum of the uncertainties of the two depths, hence approximately 95%.

Table I
Descriptive statistics of cyanide and sum of PAH's in the area, spatially stratified according to historical information

	Nr.	Mean	Std. dev	s_e	Minimum	Maximum	Skewness	Number of observations[a]			
								<T	T-Im	Im-Iv	>Iv
CN (mg kg^{-1})											
Coal-tar tank	7	8.1	6.9	4.0	3.5	16.0	1.67	4	2	1	0
Tar and ammonia tank	13	43.2	62.6	18.1	1.1	213.0	2.05	1	6	5	1
Spent oxide building	14	47.9	94.3	26.1	0.7	340.0	2.89	2	5	5	2
Gas tank	20	98.7	245.8	68.2	1.8	900.0	3.37	7	6	4	3
Outside area	88	23.4	58.8	8.4	0.1	300.0	3.67	44	33	6	5
ΣPAH (mg kg^{-1})											
Coal-tar tank	6	0.08	0.14	0.06	0.0	0.35	2.19	6	0	0	0
Tar and ammonia tank	6	1.30	3.14	1.28	0.0	7.70	2.45	5	1	0	0
Spent oxide building	2	2.25	3.18	2.25	0.0	4.50	–	1	1	0	0
Gas tank	16	28.84	77.44	19.36	0.0	304.40	3.45	9	5	1	1
Outside area	61	242.80	1766	226.13	0.0	13800	7.79	45	9	4	3

[a] T:target level.
Im: Intermediate level.
Iv: Intervention level.

Table II
Descriptive statistics according to vertical stratification on the basis of three distinguished depth layers

	Depth	Nr. of samples	Mean (mg kg^{-1})	Std.dev. (mg kg^{-1})	Median (mg kg^{-1})
CN	0–2 m	56	48.29	135.78	4.25
	2–4 m	32	22.75	62.17	1.85
	4–7 m	38	13.80	44.79	0.00
ΣPAH	0–2 m	28	29.81	70.49	3.05
	2–4 m	18	818.85	2180.04	0.12
	4–7 m	29	4.94	24.67	0.00

5. Results and Discussion

Summary statistics of collected data are given in Table I. For both *CN* and ΣPAH the mean values are above the intermediate- and below the intervention-level. So without a spatial analysis the area would be classified as moderately polluted. However, large maximum values and large coefficients of variation (3.6 mg kg^{-1} for *CN* and 6.3 mg kg^{-1} for ΣPAH), as well as the 84 and 25 observations for *CN* and ΣPAH, respectively, above the target-level, and the 11 and 4 observations above the intervention-level indicate that sub-locations are likely to be heavily polluted. This also illustrates the positively skewed distributions, which are commonly encountered in environmental studies.

The area was statified horizontally by defining polygons according to historical information, and by considering the observations within each polygon. The strata have very different mean values, for example the mean *CN*-concentration in the gas tank area was 98.7 mg kg^{-1}, whereas, quite surprisingly, the concentrations in the coal-tar tank area were very low (8.1 mg kg^{-1}). Also, the ΣPAH mean values ranged from 0.08 mg kg^{-1} within the coal-tar tank area to 242.7 mg kg^{-1} in the outside area due to a single outlier. Moreover, four out of seven *CN* observations within the coal-tar tank area were below the target value, whereas seven out of twenty were above the intermediate level in the gas tank area. The result of stratification shows that prior information is very useful for delineating areas of greater risk from those with smaller risk.

Stratification according to depth revealed a decrease in *CN*, from nearly 50 mg kg^{-1} in the top 2 m to 14 mg kg^{-1} below 4 m (Table II). For ΣPAH this decrease does not hold since there are very large concentrations at the second depth (2 - 4 m). Because of the skewness of distributions, however, the median appears to be a much better indicator of the decrease in concentrations with increasing depth (Table II).

Table III
Effects of visual observable remnants and colour on the concentrations of *CN* and Σ*PAH*

	CN			*CN*		
	No. of obs	Mean (mg kg^{-1})	St.dev. (mg kg^{-1})	No. of obs.	Mean (mg kg^{-1})	St.dev. (mg kg^{-1})
Remnant						
none	85	38.97	38.97	72	12.23	62.20
tar	6	84.32	120.11	6	2324.30	5132.30
bricks	5	201.66	349.74	4	3550.85	5918.30
coke	3	5.27	3.87	1	20.20	0.00
cyanide	4	123.75	132.26	1	0.00	0.00
coal	1	170.00	0.00	1	12.90	0.00
Colour						
normal	55	6.13	18.36	47	3.17	13.71
black	23	76.79	191.82	21	703.47	2930.86
blackblue	9	27.13	36.84	9	14.80	41.83
grey	4	31.78	51.02	1	0.02	0.00
blue	3	1.13	1.60	1	0.30	0.00
greyblue	3	2.53	3.58	2	0.24	0.14

Next, the contribution of soft information to improve estimation of the volume of polluted soil was analyzed. First, attention was focused on indirect data. For each observation it was recorded whether tar, bricks, coke, cyanide or coal were present in its immediate (within 1 m) vicinity (Table III). Samples close to bricks contain very large values of *CN* and Σ*PAH*. In the absence of any of these visual indications, measurements were relatively low (mean of *CN* and of Σ*PAH* were equal to 39 mg kg^{-1} and 12 mg kg^{-1}, respectively) although the presence of coke gave even lower values. This is an indication of some visual forms as indicators for high pollution levels. Second, attention was focused on colouring. For both *CN* and Σ*PAH* colouring is an indicator of pollution: the darker the colour, the higher are the measured concentrations. For *CN* binding to iron causes FeCn, which gives a blue colouring (Prussian Blue) to the soil, whereas Σ*PAH* is darkly coloured by itself, hence giving a darker colour to the soil as well with increasing concentrations. For a relatively small number of observations, the typical almond smell of cyanide as well as the smell of Σ*PAH* was also recorded at a safe distance and translated to a scale from 1 to 5, corresponding to no smell to very strong smell, respectively. Correlations between the first organoleptic observation (smell) and measured concentrations were usually weak (around 0.3) because smelling is disturbed by the presence of natural sulphur in the soil. Also, smell from a soil sample taken below the groundwater level can originate from the contaminated groundwater, which does not necessarily mean that the soil itself is contaminated. The

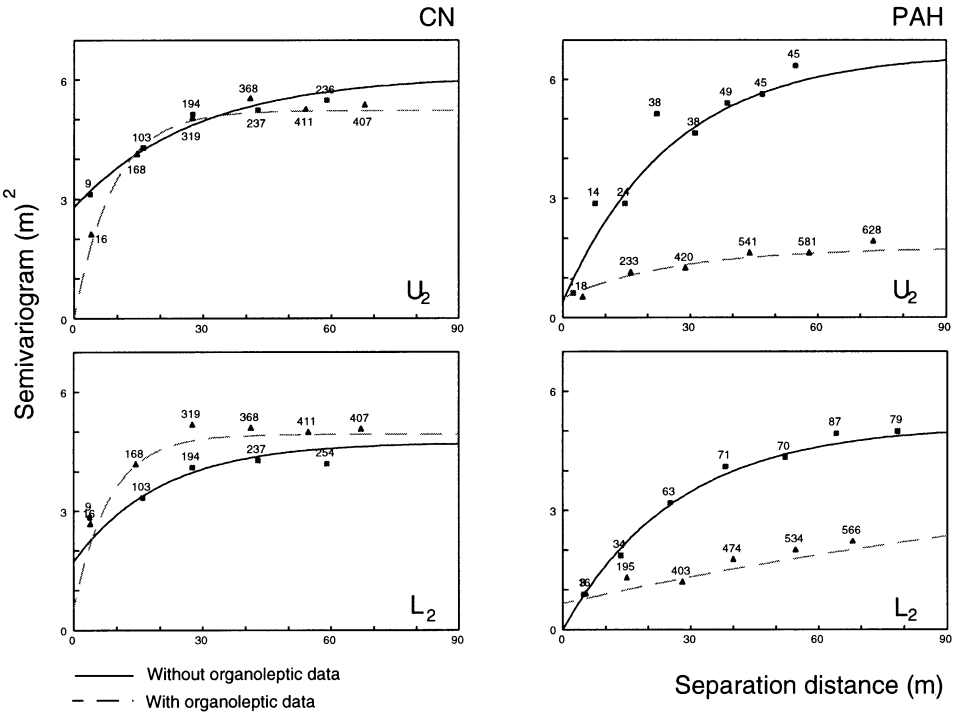

Figure 3a. Variograms for the depth to the upper surface of the first layer (U_1) and the depth to the lower surface of the first layer (L_1) for both CN and ΣPAH.

correlation coefficient between the first and the second organoleptic observation was equal to 0.921, probably caused by simultaneous collection of both data. This implies in this study, that organoleptic observations as such, either visual or by smell, are unreliable to determine the pollution level quantitatively.

As a third step additional observations on the starting and the final depth of polluted depth intervals were derived from the organoleptic observations. Estimated variograms for U_1, L_1, U_2, and L_2, for both *CN* and Σ*PAH* are presented in Figures 3a and 3b. For observations without organoleptic information, a relatively large nugget effect was obtained for *CN*, ranging from 1/3 to 1/2 of the sill value. This nugget effect almost disappears when depths based on organoleptic observations are included because of the larger number of observations. However, the sill value remains the same, indicating that the total variability is not reduced when organoleptic observations are included. In contrast, the size of the nugget effect for Σ*PAH* was negligible. Accounting for organoleptic Σ*PAH* observations reduces substantially the sill value, often by more than 50%.

Predictions by kriging were made along four transects, which cross the study area. Contamination with *CN* is shown in Figure 4; a similar figure was derived for Σ*PAH*. Both the depths where the critical target value were exceeded as well

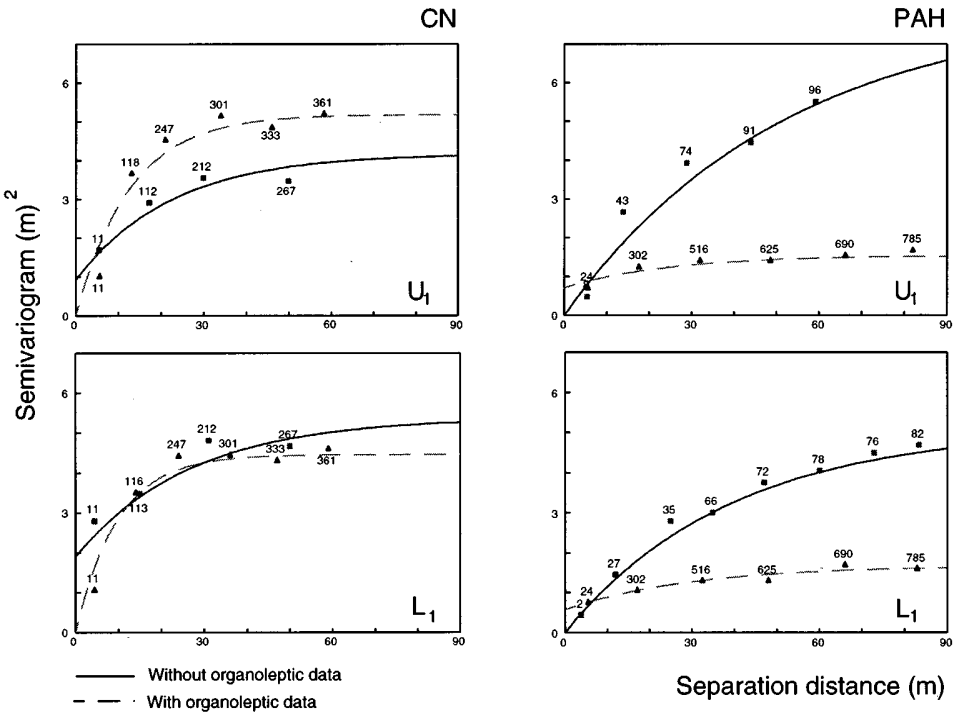

Figure 3b. Variograms for the depth to the upper surface of the second layer (U_2) and the depth to the lower surface of the first layer (L_2) for both CN and ΣPAH.

as the associated uncertainties are displayed. Since negative depths are unrealistic, negative upper boundaries were set equal to zero. Close to buildings the target value for both *CN* and Σ*PAH* is exceeded just below the soil surface. The uncertainty increases with increasing distance to the buildings.

Next, the volume of contaminated soil was determined (Table IV). The volume of soil contaminated by *CN* ranges from 55500 m^3 for the target value to 7700 m^3 for the intervention level without using any soft data. The uncertainty associated with these values is quite large, due to the large spatial variability of the contaminant and the relatively small number of observation points. Similar values were obtained for Σ*PAH*. Of interest as well is the volume of moderately polluted soil between the target and the intermediate level, which is estimated to be equal to 30700 m^3. If soft data are used as well, the volume between these levels increases to nearly 70000 m^3. We also notice that the uncertainty of the volume above the target threshold reduces by 4% for *CN* and for Σ*PAH* by almost 16%. A reduction is expected because of the increasing number of observations. The apparent difference between the two variables is due to a large reduction in sill value of the variogram of Σ*PAH* when using organoleptic observations. Finally, the two volumes were overlayed using a GIS and the amount of soil where either of the two contaminants exceeds the

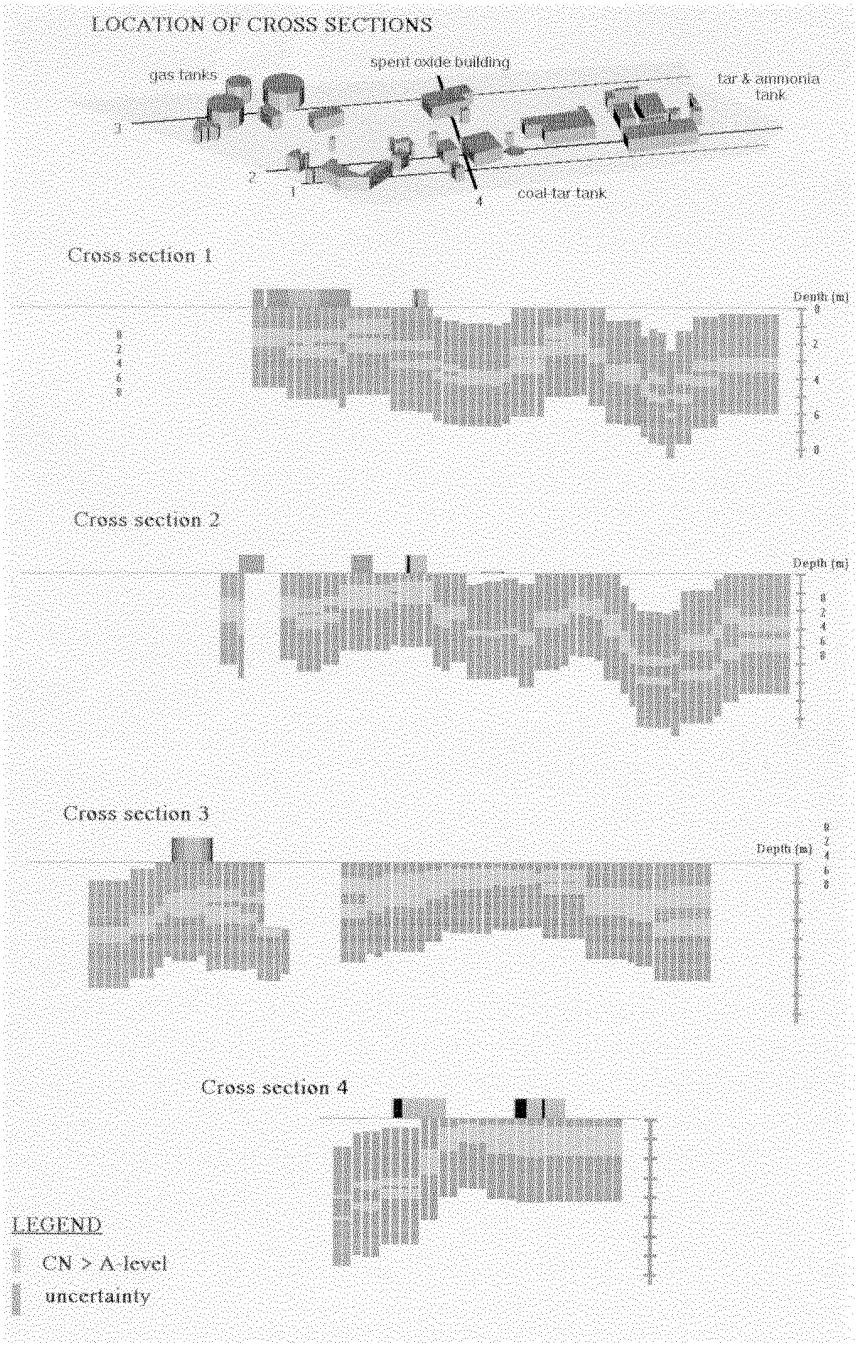

Figure 4. Polluted depth intervals of soil predicted by kriging and the associated uncertainty.

Table IV

Predicted volumes (V) of contaminated soil (m^3) and the associated uncertainties (Δ) for *CN* and ΣPAH, for different classes of contamination

	CN				ΣPAH			
	Without soft data		With soft data		Without soft data		With soft data	
Level[a]	V	Δ	V	Δ	V	Δ	V	Δ
<T	280500		267200		312700		348900	
>T	55500	191900	68800	183900	23300	205200	17100	171800
>Im	24800	211200			6600	228400		
>Iv	7700	226700			3200	270700		
T-Im	30700				16700			
Im-Iv	17100				3400			

[a] T: Target level.
IM: Intermediate level.
Iv: Intervantion level.

Table V

Predicted volume (V) (m^3) of contaminated soil as well as the associated uncertainties (Δ)

Contaminant		Without soft data	With soft data
CN	V	55500	68800
	Δ	0–256400	0–252700
ΣPAH	V	23300	17100
	Δ	0–228500	0–188900
CN and ΣPAH	V	69900	74900
	Δ	0–221200	0–184500

critical target level was determined (Table V). Also for the total volume we notice a substantial reduction in the uncertainty, with approximately 20% when organoleptic observations are included.

6. Conclusions

The use of soft data, in the sense defined in this paper, appeared to be feasible within a GIS context. Since they can be expressed as a combination of observations and coordinates, they could be used within any GIS, allowing the application of standard operations as well as geostatistics. A quantitative translation of organoleptic observations to concentration of pollutants in the soil was not useful in this study because of the low correlation between these two measurements. However, *indirect* use of organoleptic data to calculate depths at which threshold values are exceeded was much more beneficial.

Regarding the quality of the data, sampling is biased for two reasons: severely polluted key areas were not sampled, because they are known to be polluted, while the outside area was neglected. Hence, relatively few observations are far below the target level, while also the number of observations above the intermediate level is relatively small. Since the current study aims to use soft data to determine properly the volume of soil with measurements above the target value, bias of the sampling strategy is an advantage, in that very large observations are less informative than observations close the critical level. It is not appropriate, though, to use the same data for many other purposes, such as determining the volume of soil that is above the intermediate level or the intervention level.

In this study a distinction has been made between three different forms of soft data, all identified with organoleptic observations. However, the practice of soil sanitation activities recognizes another, possibly very interesting, type of soft, organoleptic, data. Any experienced surveyor may be able to assess qualitatively the hydrological conditions, in particular at a local scale. Until now such information is difficult to model within a GIS. A better relation between hydrological models acting at a point scale and GIS should be established, allowing inclusion of soft models. This would require within a GIS a direction oriented approach for data: soil moisture and groundwater have a preferential flow direction, and a raster GIS should allow the inclusion of a cause-effect relation for data. The degree of uncertainty as modelled with the current approach is substantial, in particular from a cost-decision making point of view. A reduction of these costs could be expected by accounting for soft data within a GIS.

Acknowledgements

This study was carried out with grants from the provincial board of Overijssel and from the Netherlands Integrated Soil Research Program, PCBB. The manuscript has benefitted considerably by critical and constructive comments of M. Molenaar at the Dept. of Land Surveying and Remote Sensing of the Wageningen Agricultural University.

References

Burrough, P. A.: 1989, *Journal of Soil Science* **40**, 477.
Burrough, P. A.: 1992, *International Journal of Geographical Information Systems* **6**, 1.
Cressie, N. A. C.: 1991, *Statistics for Spatial Data*, Wiley, New York.
Dowd, P. A.: 1993, 'Geological and Structural Control in Kriging', in: *Geostatistics Tróia '92*, Kluwer, Acad. Publ., Dordrecht, 1993, pp. 923–935.
Journel, A. G.: 1986, *Mathematical Geology* **18**, 269.
Leenaers, H., Burrough, P. A. and Okx, J. P.: 1989, 'Efficient Mapping of Heavy Metal Pollution on Floodplains by Co-Kriging from Elevation Data', *Three-Dimensional Applications in Geographic Information Systems*, Taylor and Francis, London, 1989, pp. 37–51.
Matheron, G.: 1965, *Les variables régionalisées et leur estimation*, Edn. Masson, Paris.

Molenaar, M.: 1991, *Journal of Photogrammetry and Remote Sensing* **46**, 85.
Raper, J. F. and Kelk, B.: 1991, in *Geographical Information Systems*, Principles and Applications. Longman, London, pp. 299–311.
Rikkers, R., Molenaar, M. and Stuiver, J.: 1994, *International Journal of Geographical Information Systems* **8**, 243.
Stein, A.: 1994, *Geoderma* **62**, 199.
Stein, A., Staritsky, I. G., Bouma, J. and Van Groenigen, J. W.: 1995, *International Journal of Geographical Information Systems* **9**, 5.
Staritsky, I. G., Sloot, P. H. M. and Stein. A.: 1992, *Water, Air, and Soil Pollut.* **61**, 1.
VROM: 1990, *Guidelines for soil sanitation*. Report nr. 01D8428, Dutch Ministry of the Environment, The Hague (in Dutch).
Webster, R. and Oliver, M. A.: 1990; *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, New York.
Zhu, H. and Journel, A. G.: 1993, 'Formatting and Integrating Soft Data: Stochastic Imaging Via the Markov-Bayes Algorithm', in *Geostatistics Tróia '92*, Kluwer, Acad. Publ. Dordrecht, 1993, pp. 1–12.