Geologie en Mijnbouw / Netherlands Journal of Geosciences 79 (4): 441-447 (2000)

Designing efficient sampling schemes for reconnaissance surveys of contaminated bed sediments in water courses

D.J. Brus[1], M.J.W. Jansen[2] & W.F. de Haan[3]

[1] Corresponding author; Alterra, Green World Research, P.O. Box 47, 6700 AA, WAGENINGEN, the Netherlands; e-mail: D.J.Brus@Alterra.wag-ur.nl
[2] Centre for Biometry, Plant Research International, P.O. Box 16, 6700 AA, WAGENINGEN, the Netherlands; e-mail: m.j.w.jansen@plant.wag-ur.nl
[3] Witteveen & Bos Consulting Engineers, P.O. Box 233, 7400 AE, DEVENTER, the Netherlands; e-mail: W.dHaan@witbo.nl

Manuscript received: 28 January 1999; accepted in revised form: 15 June 2000

Abstract

A method for designing efficient sampling schemes for reconnaissance surveys of contaminated bed sediments in water courses is presented. The method can be used in networks of water courses, for instance to estimate the total volume of bed sediment of a defined quality class. The water courses must be digitised as arcs in a Geographical Information System.

The method comprises six steps: (1) stratifying the water courses; (2) choosing a variogram; (3) calculating the parameters of the variance model; (4) choosing a compositing scheme; (5) choosing the values for the cost-model parameters; and (6) optimising the sampling scheme. The method is demonstrated with a survey of the main water courses in the reclaimed areas of Oostelijk Flevoland and Zuidelijk Flevoland.

Keywords: bed sediments, dredging, geostatistics, sampling, simulated annealing, soil survey

Introduction

The bed sediments in most water courses in the Netherlands are contaminated to some extent, and processing of dredged sediment is regulated by law. A system has been developed for classifying dredged sediment on the basis of its chemical quality. Dredged sediment of quality classes three and four is considered to be highly contaminated and has to be removed to confined disposal sites or purposely prepared treatment plants. The quality of the bed sediments to be dredged has therefore to be assessed before dredging begins. This can be done by sampling according to protocols for reconnaissance surveys and further investigations (Lamé & Bosman, 1993). Application of these protocols on a regional scale requires large numbers of samples, however. Moreover, sampling according to these protocols does not provide insight into the accuracy and precision of the results.

A geostatistical method, FAST (Fouten Analyse Saneringstraject, Error Analysis for the Remediation phase), was recently developed. It can be used to optimise intensive sampling schemes for detailed mapping of the contaminated sediment (RIZA, 1995). A reconnaissance survey of the chemical quality generally suffices at the start of a sediment-quality study. The sample size required for such reconnaissance surveys is much smaller than for detailed mapping surveys. Detailed maps of the most contaminated parts of the network may be required during the remediation phase. A method has been developed for designing efficient sampling schemes for reconnaissance surveys of sediment quality in water courses, with the objective to optimise the first phase. This new method is based on a general method for designing soil-survey schemes developed by Alterra (Domburg et al., 1994).

https://doi.org/10.1017/S0016774600021934 Published online by Cambridge University Press

General description of a case study

The method is illustrated by a survey of the sediment quality in the main water courses of the polders of Oostelijk en Zuidelijk Flevoland, in the centre of the Netherlands. A sampling scheme was designed for this area to estimate the total volume of bed sediment of quality class ≥ 2. This quantity can be calculated by estimating the mean thickness of the sediments of this quality, and multiplying this by the area of the water courses. To define the thickness of the bed sediment of quality class ≥ 2 at a specific point, it is assumed that the sediment layer has been sampled completely and mixed thoroughly. If, after mixing, the quality class appears to be < 2, the thickness is taken to be 0; if the quality class is ≥ 2, the thickness equals the total thickness of the bed sediment. The target variable can then be expressed as the product of the total thickness and a quality indicator that has a value of 1 if the quality class is ≥ 2, and a value 0 if the quality class is < 2.

A digital representation of the water courses must be made before designing the sampling scheme, because the sampling variance and the costs of a sampling scheme are determined partly by the size and geometry of the study area. Moreover, the digital representation can be used as a sampling frame for drawing a sample after an optimum sampling design has been established. The procedure described below uses ARC-INFO files. The water courses must be digitised as 'arcs', not as polygons. The width of the water courses is an attribute of the arcs. If a water course widens at a certain point, it has to be split in two. The ARC-INFO file should be carefully checked for dangling nodes, which must be removed at bridges and other places where the water courses are connected in reality. The main water courses in the two Flevopolders are shown in Figure 1.

Method

The method followed comprises the following steps:
1. stratification of the water courses;
2. choice of a variogram;

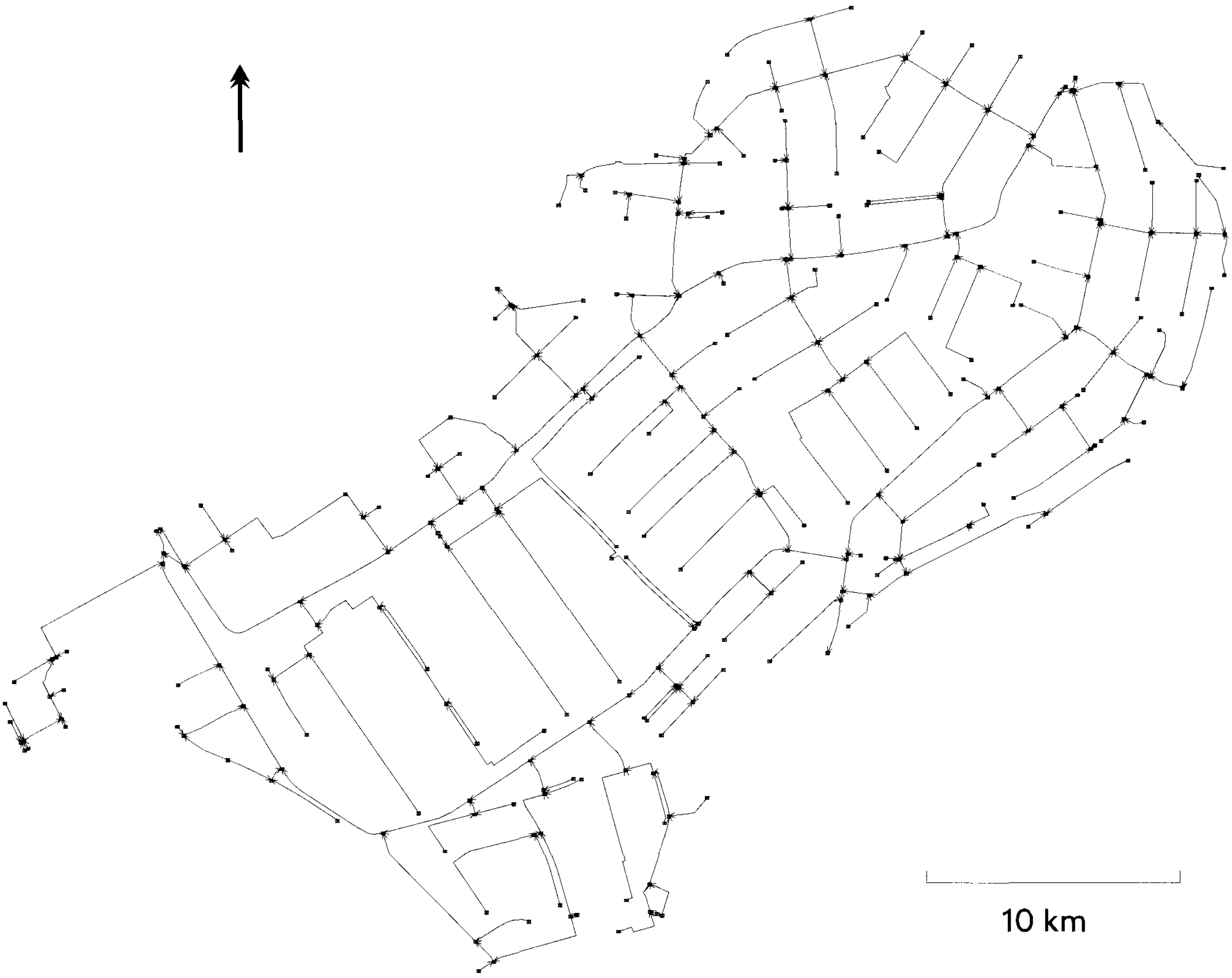

Fig. 1. The main water courses (so-called 'vaarten' and 'tochten') in Oostelijk Flevoland and Zuidelijk Flevoland.

https://doi.org/10.1017/S0016774600021934 Published online by Cambridge University Press

3. calculation of the parameters of the variance model;
4. choice of a compositing scheme;
5. choice of values for the cost-model parameters;
6. optimisation of the sampling scheme.

These steps are explained in more detail in the following subsections.

Stratification of the water courses

The precision of sampling schemes can, as a rule, be increased by dividing the study area into several more or less homogeneous subregions, referred to as 'strata'. Soil maps and land-use maps can be used for the purpose if the contamination is considered to be related to soil type and land use (see, for instance, Brus, 1994). If no prior information is available on the spatial variation of the soil property, a geographical stratification can be applied.

The water courses of Flevoland were grouped into two geographical strata: those in Oostelijk Flevoland and those in Zuidelijk Flevoland.

Choice of a variogram

The sampling variance of a survey scheme has to be predicted before sampling. This can be done using a variance model. A variogram is required for the calculation of the model parameters. Measurements from the area under study should preferably be used to estimate these variograms. If there are no such measurements, however, measurements from other similar areas can be used.

We adopted an isotropic, double spherical variogram (without nugget effect) for errorless measurements of the thickness of bed sediments of quality class ≥2 for the Flevopolders. The range and sill of the short-range variogram are 35 m and 450 cm^2, respectively; the range and sill of the long-range variogram are 11000 m and 75 cm^2, respectively. The width of the transects varies from 5 m to 50 m, so that the contribution of the long-range spherical variogram to the spatial variation within transects is negligible. The variogram is based on (1) echo soundings in the two largest water courses of the Flevopolders (Hoge Vaart and Lage Vaart), which give a rough idea of the spatial variation of the thickness of the bed sediment (regardless of its quality), (2) chemical analysis of 62 composite samples from the Flevopolders, and (3) the results of a study on the spatial variation of chemical properties in the Langbroekerwetering, several dozens of kilometres to the west of the Flevopolders (Van der Perk, 1996).

Calculation of the parameters of the variance model

The sampling variance is not only determined by the number of points (sample size) but also by the sampling design (selection procedure). The points are selected in three stages (Fig. 2). In the first stage, a water course (arc) is drawn independently n_h times from stratum h , with probabilities proportional to the area of the water course and with replacement. In the second stage, any time a water course is drawn, m_h transects perpendicular to the direction of the axis of the water course are selected from this water course by simple random sampling (*SI*), that is with equal probability and independently. If a water course is selected r times in the first stage, m_h transects are selected r times from this water course. In the third stage, k_h points from each selected transect are selected by *SI*.

In technical terms, the water courses are the primary sampling units (*psu's*), the transects the secondary sampling units (*ssu's*), and the points the tertiary sampling units (*tsu's*).

If m_h and k_h are larger than 1, this selection procedure is a stratified three-stage sampling design (Cochran, 1977). If either m_h or k_h equals 1, a stratified two-stage design is obtained, and if both m_h and k_h equal 1, a stratified simple random sampling design is

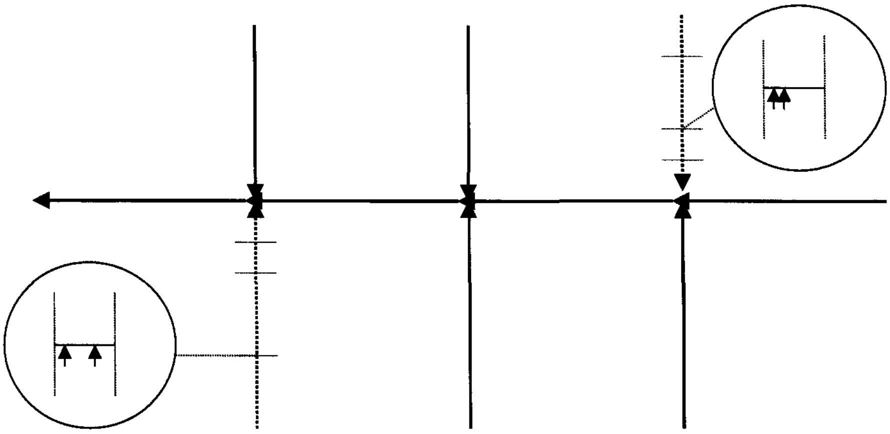

Fig. 2. Notional example of three-stage sampling from a network of water courses. In the first stage, two watercourses are selected (with probabilities proportional to their area and with replacement); in the second stage, three transects per water course (with equal probability); and in the third stage, two points per transect are selected (with equal probability).

https://doi.org/10.1017/S0016774600021934 Published online by Cambridge University Press

obtained. It is possible to set h equal to 1 (no stratification); simple (unstratified) counterparts of the designs are obtained in such a case. There is no need to fix the parameters, m_h and k_h, in advance at the value of 1, because this will follow from the optimisation if a two-stage or a one-stage design is more efficient. Note that m_h and k_h are constant for all water courses in stratum h.

For the stratified three-stage design, the sampling variance, $\hat{V}$, can be predicted by:

$$\hat{V} = \sum_{h=1}^{L} \left[\frac{A_h}{A}\right]^2 \sum_{i=1}^{N_h} \frac{A_{ih}}{A_h} \left[\frac{\overline{\gamma}_h - \overline{\gamma}_{psu_{ih}}}{n_h} + \frac{\overline{\gamma}_{psu_{ih}} - \overline{\gamma}_{ssu_{ih}}}{n_h m_h} + \frac{\overline{\gamma}_{ssu_{ih}}}{n_h m_h k_h}\right] \qquad \text{(Eq. 1)}$$

in which equation:

L is the number of strata;

A is the area of all water courses in the study area;

n_h is the total number of water courses (primary units) in stratum h;

A_{ih} is the area of the ith water courses in stratum h;

A_h is the total area of the water course in stratum h;

$\overline{\gamma}_h$ is the mean semivariance of all pairs of points in stratum h;

$\overline{\gamma}_{psu,\,ih}$ is the mean semivariance of all pairs of points in the ith water course (psu) of stratum h;

$\overline{\gamma}_{ssu,\,ih}$ is the mean semivariance of all pairs of points in a transect (ssu) in the ith water course (psu) of stratum h.

Equation (1) can be used for all sampling design types mentioned above. The three numerators of Equation (1) are predictions of spatial variances, i.e. the spatial variance between the mean thickness of water courses in stratum h, the spatial variance between the mean thickness of transects in watercourses of stratum h, and the spatial variance of the thickness at points in transects of stratum h.

The mean semivariance of all pairs of points within transects and the mean semivariance of all pairs of points within water courses of a stratum are approximated by the one-dimensional and two-dimensional Cauchy-Gauss method, respectively (Journel & Huijbregts, 1978). We assume that the thicknesses of the bed sediment at two points within different water courses are independent. Regarding the dimension of the network this implies that the mean semivariance of all pairs of points within a stratum can be approximated by the sill of the variogram. Table 1 shows the values of the parameters of the variance model (Eq. 1).

Table 1. Calculated values for the parameters of the variance model (numerators of Eq. 1).

parameter	Oostelijk Flevoland	Zuidelijk Flevoland
$\overline{\gamma}_h - \sum_{i=1}^{N_h} \frac{A_{ih}}{A_h} \overline{\gamma}_{psu_{ih}}$	56	56
$\sum_{i=1}^{N_h} \frac{A_{ih}}{A_h} (\overline{\gamma}_{psu_{ih}} - \overline{\gamma}_{ssu_{ih}})$	319	318
$\sum_{i=1}^{N_h} \frac{A_{ih}}{A_h} \ \overline{\gamma}_{ssu_{ih}})$	165	166

Choice of a compositing scheme

The variance of Eqation (1) is the variance of the sampling error only. Besides the sampling error we have also a measurement error. The contribution of this error to the total error depends on the composite grouping scheme. When the costs of laboratory analysis are considerable, budget can be saved by grouping samples taken at points into composite samples. The more samples are grouped, the more budget is saved; on the other hand, the contribution of the measurement error to the total error increases. There are three possibilities: (1) compositing at the level of the primary units, i.e. all samples within a water course are grouped into one composite; (2) compositing at the level of the secondary units; (3) no compositing. For compositing at the level of the primary units, the variance of the measurement error must be added to the numerator of the first term in Equation (1); at the level of the secondary units it must be added to the numerator of the second term, and it must be added to the numerator of the third term if samples are not grouped into composites.

For the Fleverwaard we designed a sampling scheme for compositing at the level of the primary units. The variance of the measurement error was taken to be 25 cm^2.

Choice of the values for the cost-model parameters

The total cost equals the sum of the costs per stratum. The cost per stratum, C_h, is predicted by:

$$C_h = c_s t_h + c_e t_h + s_h c_l \qquad \text{(Eq. 2)}$$

in which formula:

c_s is the survey costs per hour;

c_e is the equipment costs per hour;

c_l is the costs of laboratory analysis per sample;

t_h is the time needed for fieldwork in stratum h;

s_h is the number of (composite) samples from stratum h to be analysed in laboratory.

The time for fieldwork is calculated as the sum of the access time (the time needed for travelling to the sampling units and locating the sampling units) and the observation time (time spent at the sampling points for observation and sampling).

The access time is calculated as the sum of the time to access the selected water courses (*psu's*), the time to access the selected transects (*ssu's*) within these water courses, and the time to access the selected points (*tsu's*) within these transects. Note that, because water courses are selected with replacement, the number of water courses in the sample can be less than the number of primary unit draws. The expected number of different water courses in the sample per stratum, $E[n_h{}^{diff}]$, equals:

$$E\,[n_h^{diff}] = \sum_{i=1}^{N_h} \left\{ 1 - \left(1 - \frac{A_{ih}}{A_h} \right)^{n_h} \right\} \qquad \text{(Eq. 3)}$$

The total access time of water courses per stratum, t_{a1h}, can then be calculated with:

$$t_{a1h} = \tau_{a1}\, E[n_h^{diff}] \qquad \text{(Eq. 4)}$$

in which formula τ_{a1} is the mean access time per watercourse. This mean access time per water course is largely determined by the time needed for loading and unloading the boat. We assume that this happens every time another water course is sampled. The access time of secondary units is simply calculated as the product of the mean access time per secondary unit, τ_{a2}, and the number of selected secondary units in stratum h:

$$t_{a2h} = \tau_{a2} n_h m_h \qquad \text{(Eq. 5)}$$

Similarly, the access time for tertiary units is calculated by:

$$t_{a3h} = \tau_{a3} n_h m_h k_h \qquad \text{(Eq. 6)}$$

The observation time is calculated as the product of the mean observation time per point (τ_o) and the number of selected points (*tsu's*).

For the Fleverwaard, the number of samples to be analysed, s_h, equals the number of primary unit draws, n_h. Note that, if a water course is selected more than once, more than one composite sample is taken from this water course and analysed in the laboratory. The values of the cost parameters used for the Flevopolders are presented in Table 2.

Table 2. Chosen values for the parameters of cost model.

cost-model parameter	value
c_s (NFL)	100
c_e (NFL)	50
c_l (NFL)	500
τ_{a1} (h)	1.5
τ_{a2} (h)	0.1
τ_{a3} (h)	0.01
τ_o (h)	0.25

Optimisation of the sampling scheme

The more transects per water course and the more points per transect selected, the more the sampling points will cluster in the network. This will generally lower the costs of sampling, but may raise the sampling variance. After a budget has been established, the sampling variance is minimised under the constraint that the costs do not exceed the budget. A small excess in the budget is permitted, if this will decrease the variance more than on average. This is achieved with an objective function with two terms:

$$\text{Max}(\frac{1}{V} - \lambda \beta) \qquad \text{(Eq. 7)}$$

in which formula:

V is the variance;

λ is a weighting parameter;

β is a penalty function (the penalty is zero for costs far below the budget; when the costs approach the budget, the penalty increases slightly, but the penalty increases strongly as soon as the budget is reached).

For the penalty we chose the following function:

$$\beta(\delta;\alpha) = \alpha\ [\ln(1 + e^{\delta \ln(2)/\alpha})/\ln(2)] \qquad \text{(Eq. 8)}$$

in which formula:

δ is the difference between the predicted costs and the budget;

α is a smoothing parameter.

This function has two asymptotes: for $\delta \rightarrow -\infty$ the X-axis ($b = 0$) and for $\delta \rightarrow \infty$ the 45-degree line ($\beta = \alpha$). At the point $\alpha = 0$, the function has value $\beta(0;\alpha) = \alpha$; while the derivative equals $\beta(0;\alpha) = 0.5$. The smoothness of the function is determined by the parameter α. In the limiting case of $\alpha = 0$, the derivative is discontinuous; the larger α, the smoother the curve connecting the two asymptotes. Figure 3 shows this function for $\alpha = 100$, 500 and 1000. The parameter, λ, of the objective function must be greater than the extra information per money unit near the budget. For smaller values, the problem of Equation (7) has no optimum solution.

The objective function is maximised by simulated annealing (Kirkpatrick et al., 1983). This is a heuristic, random search technique. The search is iterative, starting with randomly chosen values for the numbers of sampling units. For L strata, $3L$ numbers are to be

https://doi.org/10.1017/S0016774600021934 Published online by Cambridge University Press

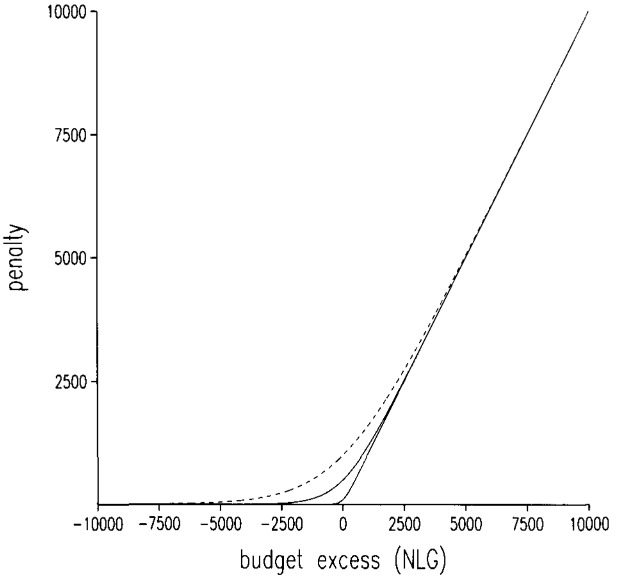

Fig. 3. The penalty function, β (εδ;α), for α = 100, 500 and 1000.

optimised. A slight random change of these numbers is 'proposed' in each iteration step. The proposal is always accepted if the value of the objective function increases. In the early stages of the iterative process, however, changes that lead to a decrease of the objective function can be accepted with a positive probability. This policy reduces the probability of finding a sub-optimum solution that cannot be improved by a small modification of the design parameters, although a much better solution exists further away. The probability of accepting worse solutions decreases to zero during the iterative process.

Two strata were defined for the Flevopolders, so that six numbers had to be optimised. Table 3 shows the optimum sample sizes for ten budgets. The optimum number of transects per water course is 8, with some exceptions. The optimum number of points per transect amounts to one in all cases.

Table 3. Optimum numbers of water courses (*n*), transects per water course (*m*) and points per transect (*k*) for a budget of NFL 10,000 -100,000.

budget (NFL)	Oostelijk Flevoland			Zuidelijk Flevoland		
	n	*m*	*k*	*n*	*m*	*k*
10,000	5	8	1	4	6	1
20,000	10	9	1	7	7	1
30,000	16	8	1	10	8	1
40,000	21	8	1	13	9	1
50,000	26	8	1	17	9	1
60,000	30	9	1	21	8	1
70,000	37	8	1	25	8	1
80,000	43	8	1	28	8	1
90,000	48	8	1	32	8	1
100,000	54	8	1	35	8	1

Discussion and conclusions

Some uncertainty about the variograms and the values of the cost model parameters always exists. As the optimisation result depends on the values of the model parameters, the optimum sampling design is also uncertain. The degree of uncertainty depends, however, on how sensitive the optimisation result is to changes in the model parameters. The optimisation step should consequently be repeated with changed values for the model parameters. If the optimisation result differs greatly, additional information on the model parameters should be collected.

In addition to the optimum numbers of sampling units, the method also provides predictions of the costs and the variance. By calculating the costs and variance for a range of budgets, information is obtained that can be used to decide on the ultimate budget (Fig. 4).

The method is also appropriate for the design of efficient schemes aiming at estimating the volume of (contaminated) bed sediment in several subregions. The number of subregions must be adjusted to the number of points that can be sampled with the budget.

The restriction that the numbers of selected transects per water course, and of selected points per transect are constant for all water courses within a stratum may lead to unnecessarily high sampling variances if the length of the water courses and the width of the transects varies considerable within the stratum. The network should, if possible, be split up into water courses (primary units) of approximately equal length. If the width of these primary units varies considerable, one might decide to calculate optimum numbers of points per transect per water course or group of water courses with approximately equal width.

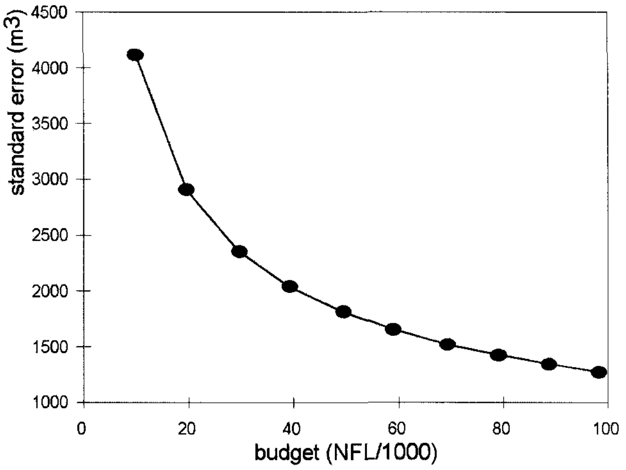

Fig. 4. Predicted standard deviation of estimated volume of contaminated bed sediment as a function of the budget.

https://doi.org/10.1017/S0016774600021934 Published online by Cambridge University Press

Note that grouping the samples into composite samples leads to unbiased estimates of the volume of contaminated bed sediment only if the thickness (regardless of quality) and the quality indicator are independent. If there are serious doubts on this, one might decide to analyse the samples separately (no compositing). Another reason for not grouping samples into composites is that the measurements can then be used in a subsequent inventory for mapping the bed-sediment quality by spatial interpolation. If samples are not grouped into composites, the optimum number of transects per water course decreases to 3 (lowest budgets) or 1 (highest budgets). The optimum number of points per transect is then 1 again . In this case, strong clustering of points in a restricted number of water courses is less efficient because this reduces the costs only marginally (the costs are largely determined by the laboratory costs), whereas strong clustering leads to a considerable increase of the variance.

Acknowledgements

We acknowledge Jaap De Gruijter (Alterra) for his statistical advice, and Martin Veul (Witteveen & Bos) for stimulating us to apply theory in tackling practical problems.

References

Brus, D.J., 1994. Improving design-based estimation of spatial means by soil map stratification; a case study of phosphate saturation. Geoderma 62: 233-246.

Brus, D.J. & Jansen, M.J.W., 1998. Gestructureerd ontwerpen van efficiënte plannen voor de inventarisatie van de bodemkwaliteit in watergangen geïllustreerd met de Fleverwaard. Internal Report DLO-Staring Centrum (Wageningen) 587: 58 pp.

Cochran, W.G., 1977. Sampling techniques (3^{rd} ed.). John Wiley and Sons (New York): 428 pp.

Domburg, P., De Gruijter, J.J. & Brus, D.J., 1994. A structured approach to designing soil survey schemes with prediction of sampling error from variograms. Geoderma 62: 151-164

Journel, A.G. & Huijbregts, Ch.J., 1978. Mining geostatistics. Academic Press (London): 600 pp.

Kirkpatrick, S., Gelatt, C.D. & Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220: 671-680.

Lamé, F.P.J. & Bosman, R., 1993. Protocol voor het oriënterend onderzoek naar de aard en concentratie van verontreinigende stoffen en de plaats van voorkomen van bodemverontreiniging. SDU ('s-Gravenhage): 80 pp.

RIZA, 1995. POSW, Fact sheet 3. RIZA (Lelystad): 4 pp.

Van der Perk, M., 1996. Muddy waters: uncertainty issues on modelling the influence of bed sediments on water composition. Netherlands Geographical Studies 200: 1-190.