International Congress on Environmental Modelling and Software

5th International Congress on Environmental Modelling and Software - Ottawa, Ontario, Canada - July 2010

Jul 1st, 12:00 AM

A Study of Variance Estimators for Material Sampling Using Computerized Models of Contaminant Heterogeneity in Soil Stockpiles.

B. Geelhoed

F. P. J. Lamé

International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

A Study of Variance Estimators for Material Sampling Using Computerized Models of Contaminant Heterogeneity in Soil Stockpiles.

B. Geelhoed[a)], **F.P.J. Lamé**[b)]
a) Delft University of Technology, Mekelweg 15, 2629 JB Delft, The Netherlands, B.Geelhoed@tudelft.nl
b) Deltares, PO Box 85467, 3508 AL Utrecht, The Netherlands, frank.lame@deltares.nl

Abstract: During the sampling of contaminated soil, sampling errors are unavoidable because of the spatial heterogeneity of the contaminant distribution. The variance is a convenient indicator for the potential magnitude of these errors. Four variance estimators are constructed for use in material sampling, all of which take account of the heterogeneity and the sampling design. Based on large scale three-dimensional computerized models of contaminant heterogeneity in soil stockpiles, these variance estimators are compared using a Monte Carlo simulation of different sampling designs. The Mean Squared Error (MSE) of each variance estimator is used to assess (and compare) the performance of each variance estimator: the lower the mean square error, the better its performance.

Key words: Sampling; soil; variance; estimator; heterogeneity.

1. INTRODUCTION

1.1. Sampling Error

In order to determine if a potentially harmful contaminant in a batch of soil actually poses a threat, a first step in its risk assessment will be to analyze a sample obtained from that batch. The contaminants present will often be expressed as a concentration; e.g. mg/kg. As a series of steps (including sampling, sub-sampling and analyses) is necessary to determine the contaminants concentration, errors occurring during these steps will most likely result in a difference between the "true value" (the real mean concentration in the population) and the estimate (the concentration determined in the sample). This article will only focus on the error caused by primary sampling. Errors occurring during the following steps of sub-sampling and analysis are not considered. The sampling error (e_s) can be expressed as:

$$e_s = c_{sample} - c_{batch} \qquad \textbf{(1)}$$

where:
c_{sample} = the concentration of the contaminant in the sample.
c_{batch} = the real mean concentration of the contaminant in the batch (unknown).

1.2 Variance

The sampling error is generally expected to differ from sample-to-sample in a random way as c_{sample} of sample i will differ from c_{sample} of sample j, both randomly differing from c_{batch}. A useful quantity is therefore the sample-to-sample variance of the sampling error, denoted by the symbol $V(e_s)$. Taking the variance operator on both sides of (1) results in $V(e_s) = V(c_{sample})$, where $V(c_{sample})$ denotes the sample-to-sample variance of c_{sample}.

The sample-to-sample variance can (in principle) be estimated by taking multiple samples. Leaving the errors caused by the necessary subsequent steps of sub-sampling and analyses out of consideration, the variance $V(c_{sample})$ can be estimated by:

$$V_{est}(c_{sample}) = \Sigma_i \, (c_{sample,i} - c_{av})^2/(N\text{-}1) \qquad \textbf{(2)}$$

where:

$V_{est}(c_{sample})$ = the estimate of $V(c_{sample})$.
Σ_i = a summation over index i.
$c_{sample,i}$ = the sample concentration in the i-th sample.
N = the number of samples
i = an index of each sample (i=1,…,N).
c_{av} = the average of the N values $c_{sample,i}$.

When two or more sampling results are available, (2) can be used to estimate the sample-to-sample variance at the scale of the sample size.

1.3 Scale and variance

Estimating the concentration of the soil batch by the concentration of a sample of a specific scale introduces variability as the sample concentration will differ randomly from the true mean concentration of the batch. Consequently, a relation exists between the scale of the sample and the variance.

Limiting the sample-to-sample variance is of major importance as in practice decisions on the risks posed by the contaminants will often be based on one, or at most a small number of samples. This raises the question how the sample-to-sample variance can be practically determined and what a cost effective sampling strategy will be, considering both the scale of the sample and the number of increments.

1.4 Increments

The new approach of this article is based on information on the scale of the increments of which the sample is composed. This information is generally not available, but it is expected that it can be made available in future at the expense of an increased effort of sampling and increased costs of analyses. In section 2, it will be demonstrated that it is in principle possible to estimate the sample-to-sample variance even when using a single sample, provided that information at the level of the increments is made available (e.g. by analyzing all increments separately).

A batch of soil can be thought of as a population of increments. If each increment in the population is assigned a unique number taken from the set of numbers ranging from 1 to N_{pop}, where N_{pop} is the total number of increments in the population, the concentration (c_{sample}) can be written as a function of the individual increment properties:

$$c_{sample} = \Sigma_i \, I_i \, m_i \, c_i \, / \, (\Sigma_j \, I_j \, m_j) \qquad \textbf{(3)}$$

where:

Σ_i = a summation over index i.

Σ_j = a summation over index j.
m_i = the mass of the i-th increment in the population.
m_j = the mass of the j-th increment in the population.
c_i = the concentration in the i-th increment in the population
c_j = the concentration in the j-th increment in the population
I_i = the indicator of the i-th increment in the population.
I_j = the indicator of the j-th increment in the population.

I_i (or I_j) is one when the i-th (or j-th) increment is part of the sample and zero otherwise. This ensures that the summations in (3) only sum over the increments in the sample.

2. Estimators for the variance

It was shown by Geelhoed [2004] that (under the conditions of a constant sample mass and constant first-order inclusion probability) the concentration in a sample (defined by (3)) can be seen as a π-expanded estimator (see e.g. Särndal et al. [1992]):

$$Y_\pi = \Sigma_i I_i y_i / \pi_i$$

where:
Y_π = the π-expanded estimator for the population total of y_i.
Σ_i = a summation over index i.
y_i = the variable of interest (here: $y_i = m_i c_i / M_{pop}$, where M_{pop} is the mass of the entire population).
i = the number of an increment in the population.
I_i = the indicator of the i-th potential increment in the population.
π_i = the first-order inclusion probability of the i-th potential increment in the population (see e.g. Särndal et al. [1992]). Because π_i appears in the denominator, the estimator Y_π can only be applied if $\pi_i > 0$.

Substituting $y_i = m_i c_i / M_{pop}$ (where M_{pop} is the mass or weight of the entire population) and assuming constant first-order inclusion probabilities $\pi_i = M_s / M_{pop}$ indeed results in $Y_\pi = c_{sample}$.

The Horvitz-Thompson estimator (V_{HT}) (see e.g. Särndal et al. [1992]) and the Sen-Yates-Grundy estimator (V_{SYG}) (Sen [1953], Yates and Grundy [1953]) are defined by the following equations:

$$V_{HT} = \Sigma_i \Sigma_j I_i I_j (1 - \pi_i \pi_j / \pi_{ij}) (y_i / \pi_i) (y_j / \pi_j)$$

and

$$V_{SYG} = -(1/2) \Sigma_i \Sigma_j I_i I_j (1- \pi_i \pi_j / \pi_{ij}) (y_i / \pi_i - y_j / \pi_j)^2$$

where:
π_{ij} = the joint inclusion probability of the i-th and j-th potential increment in the population. Because π_{ij} appears as the denominator in a fraction, the estimators V_{HT} and V_{SYG} can only be applied if $\pi_{ij} > 0$. Note the following definition (which is not standard, but which allows the above simplified equations for V_{HT} and V_{SYG}): π_{ii} exists and equals π_i.

Using the "parameter for the dependent selection of particles" [1], $C_{ij} = 1- \pi_{ij}/ (\pi_i \pi_j)$, and using $y_i = m_i c_i / M_{pop}$ results in:

$$V_{HT} = -(1/M_s^2) \Sigma_i \Sigma_j I_i I_j (C_{ij} / (1 - C_{ij})) m_i c_i m_j c_j \qquad \textbf{(4)}$$

and

$$V_{SYG} = (1/(2\ M_s^2))\ \Sigma_i \Sigma_j I_i I_j\ (C_{ij} / (1 - C_{ij}))\ (m_i c_i - m_j c_j)^2 \qquad \textbf{(5)}$$

where:

M_s = the mass of a sample

C_{ij} = the parameter for the dependent selection of particles. Because of the conditions $\pi_i > 0$ and $\pi_{ij} > 0$ for applicability of V_{SYG} and $V_{HT,}$ it follows that the above equations ((4) and (5)) can only be applied if $C_{ij} < 1$, not if $C_{ij} = 1$.

Hence, the (estimated) variance depends on the parameters M_s, C_{ij}, m_i and c_i of the increments in the sample.

It is noted here that the above two variance estimators are entirely general: they are valid for all possible sampling designs (provided $C_{ij} < 1$). Differences between sampling designs are taken into account by the parameters M_s and C_{ij}: different designs can lead to different values for these two parameters. A condition for the estimator V_{SYG} to be unbiased is that the number of increments in a sample has a zero variance. V_{HT} does not have this restriction, but both V_{HT} and V_{SYG}, as given in (4) and (5), were constructed using the assumption that the sample mass (M_s) is constant. This is indeed true when both the increment mass and the number of increments are constant.

Given the fact that V_{HT} is slightly more general than V_{SYG}, it is interesting to investigate possible improvements to V_{HT}. Inspection of (4) shows that V_{HT} is potentially sensitive to systematic errors made in the determination of c_i and c_j. However, the sample-to-sample variance $V(c_{sample})$ is not influenced by systematic errors in determination of c_i, c_j and c_{sample}. Therefore it makes sense to adapt V_{HT} as follows:

$$V_{AD1} = -(1/M_s^2)\ \Sigma_i \Sigma_j I_i I_j (C_{ij} / (1 - C_{ij}))\ m_i m_j\ (c_i - c_{sample})\ (c_j - c_{sample}) \qquad \textbf{(6)}$$

where:

V_{AD1} = the adapted variance estimator based on V_{HT}.

A special scenario is Poisson sampling (see e.g. Särndal et al. [1992]). During Poisson sampling, each increment is independently subject to a probabilistic selection process. Therefore $C_{ij} = 0$ (for unequal i and j). Substituting this value for C_{ij} in the expression for V_{SYG} leads to $V_{SYG}=0$ (which underestimates the variance which is generally non-zero). Substitution of $C_{ij}=0$ and $C_{ii}=1-M_{pop}/M_s$ in the expression for V_{AD1} yields:

$$V_{Poisson} = ((1 - M_s/M_{pop})\ /M_s^2)\ \Sigma_i\ I_i\ m_i^2\ (c_i - c_{sample})^2 \qquad \textbf{(7)}$$

Even if the used sampling design differs substantially from the Poisson sampling design, the estimator $V_{Poisson}$ may in many cases still provide a reasonable variance estimate if the spatial arrangement of increments in the population can be considered to be completely random. (7) also offers considerable computational simplification with respect to (4), (5) and (6), because (7) contains only a single summation symbol, while (4), (5), and (6) require double summations. Another advantage of $V_{Poisson}$ is that it does not depend anymore on C_{ij}: the variance can also be estimated when $C_{ij}=1$, which is not possible using V_{HT}, V_{SYG} or V_{AD1}. Moreover, (7) is intuitively easier to understand: the estimate of sample-to-sample variance equals the (weighed) within sample variance $((1/M_s)\Sigma_i\ I_i\ m_i^2\ (c_i - c_{sample})^2)$ divided by the sample mass (M_s) in order to convert a within-sample variance to a between sample variance and multiplied by a finite population correction factor ($1-M_s/M_{pop}$) in order to take into account the effect of the finite population size on the variance. However, despite all these advantages of $V_{Poisson}$, there is no guarantee that the estimate obtained with $V_{Poisson}$ is accurate, especially because $V_{Poisson}$ neglects taking into account the influence of C_{ij} on the variance. Therefore, in this article the performance of V_{HT}, V_{AD1}, V_{SYG} and $V_{Poisson}$ will be compared.

3. SIMULATION STUDY

A Monte Carlo simulation is presented, which is applied to large scale three-dimensional computerized models of contaminant heterogeneity in soil stockpiles. The procedure is:

1. An independent random sample is drawn without replacement from a virtual population and c_{sample} and V_{HT} , V_{SYG} , V_{AD1}, and $V_{Poisson}$ are recorded for this sample (calculated using (3), (4), (5), (6) and (7) respectively) wherein the sample is obtained by applying incremental sampling (see section 3.2).
2. The sample “material” is put back in the virtual population.
3. Go back to step 1 until 50,000 repetitions are performed.

3.1 Model of a Population

Thirty virtual models of contaminated soil stockpiles were constructed (Lamé et al. [2005]). These models are based on actual three dimensional concentration data from three soil stockpiles (denoted by “gas”, “dpa” and “rok”) for each of which ten virtual models were constructed. Obviously, only a small fraction of the original soil lots was analyzed. However, the data obtained were used to define, within the same statistical distribution of the observations, all increments in those soil lots, wherein by spatial simulation within the series of ten models, different degrees of large scale heterogeneity were introduced. Hence, the three series of ten models represent soil lots that, with the same statistical distribution, differ considerably in their degree of spatial distribution of highs and lows, thus their degree of large scale heterogeneity.

Here, only a brief description of the geometry of the virtual populations is given. The details of construction are described elsewhere (Lamé et al. [2005]). Each population consists of a rectangular arrangement of cubic “cells” (which represent the increments) wherein for each individual cell the concentration of the contaminant is defined. Consequently, for each of the thirty models the full population in known. The arrangement of cells in each virtual populations is similar to the arrangement displayed in Figure 1, but larger.

Each increment will not only be characterized by its mass (m_i) and concentration (c_i), but also by its spatial location within the virtual population. Here this location is denoted by the triple of indices (x,y,z). The parameter for the dependent selection of particles (C_{ij}) will depend on the spatial locations of increments i and j relative to each other.

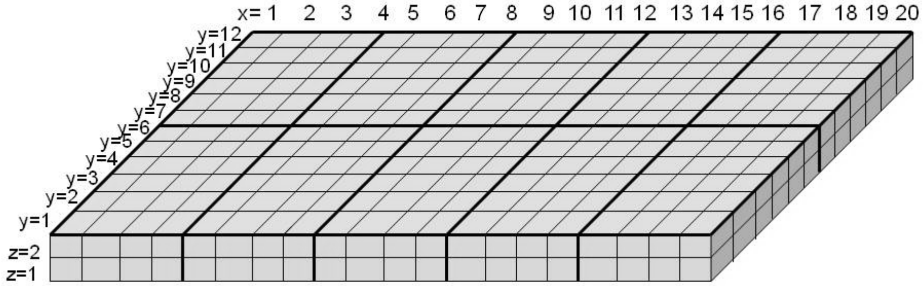


Figure 1. Model of the population. In the here-depicted model z ranges from 1 to 2; y ranges from 1 to 12; x ranges from 1 to 20; and the population is divided in 10 equally-size blocks. The models that are used here are significantly larger (and therefore more realistic).

Note: for simplicity the assumption m_i=1 is made during the simulations, which, under the boundary condition that m_i is constant, is of no effect to the results. The size of these virtual populations is as follows:

- "gas": height = 15 cells; width = 371 cells; length = 1054 cells. In this article the "gas" populations are cropped to 15x368x1052 because this makes it easier to divide the population in equally-sized blocks.
- "dpa": height = 16 cells; width = 920 cells; length = 400 cells.
- "rok": height = 16 cells; width = 500 cells; length = 735 cells. In this article the "rok" populations are cropped to 16x500x732 because this makes it easier to divide the population in equally-sized blocks.

3.2 The Sampling Design

For the purpose sampling, each "block" is subdivided in vertical stacks of increments. Each vertical stack is characterized by its location in the (x,y) plane and consists (for given x and y) of all increments with coordinates (x, y, 1), (x, y, 2), ... , (x, y, z_{max}) where z_{max} is the "height" of the virtual model (i.e. the number of horizontal layers, which was 15 or 16 depending on which virtual model was used). Hence each vertical stack contains the increments from all heights at its (x, y) location in the virtual model. If the selection of the sample only takes vertical stacks as a whole, possible large scale heterogeneity in the vertical direction will not result in biased sampling, as all heights will be equally represented in the sample. This will be the case for the here-described sampling design.

From each "block" in the population (see Figure 1) a constant number (n) of vertical stacks of increments is selected at random locations in the two-dimensional (x,y)-plane. Selection of vertical stacks of increments in a block is by simple random sampling of vertical stacks (see e.g. [3] for a precise definition of "simple random sampling"). The overall sample is formed by combining the simple random samples of vertical stacks from each block (see Figure 2 for a simple illustration).

It is noted that the here-described sampling design can be described as "stratified simple random sampling". In case the number of blocks (which equals the number of strata) equals one, this mode of sampling reduces to simple random sampling.

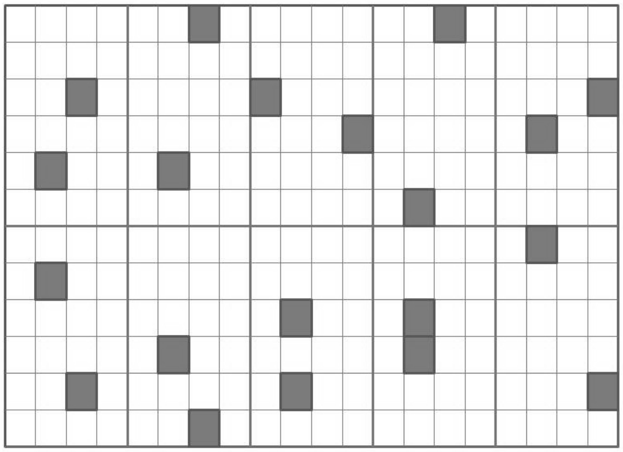


Figure 2. Schematic depiction of the sampling design. This is a top view of the same virtual population as in Figure 1. The grey squares represent the vertical stacks of increments that are part of a random sample. In the here-depicted scenario, the population is divided into 10 equally-sized blocks and from each block two vertical stacks of increments are selected

(N_{block}=10, n=2). The sample therefore consists (in this example) of 10 x 2 = 20 vertical stacks of increments.

The values of C_{ij} for the here-described sampling design are as follows:

C_{ij} = (if increment i and j belong to different blocks) 0. This follows from the fact that increments belonging to different block are selected independently.

C_{ij} = (if increment i and j belong to the same vertical stack) 1 – 1/q (where q is the first-order inclusion probability of an increment, which is equal to the ratio of the number of increments in the sample and the total number of increments in the population before the sample was taken). This value follows from the fact that for increments belonging to the same vertical stack $\pi_{ij} = \pi_i$ combined with the definition of C_{ij}.

C_{ij} = (if increment i and j belong to the same block, but to different vertical stacks) (1-q)/(n-q). A derivation is presented in Appendix A.

In a first computer experiment, using the here described sampling design, the thirty models were sampled using three different variations of incremental sampling:

1. (Simple random sampling) N_{block} = 1, n=10 (in total the sample consist of 10 stacks of 15 or 16 (depending on which virtual model was used) increments each).
2. N_{block}=4 (obtained by "quartering" the population in four equal quadrants in the (x,y) plane), n=8 (in total the sample consist of 32 stacks of 15 or 16 (depending on which virtual model was used) increments each).
3. N_{block}=16 (obtained by quartering the population and then quartering each quarter in a similar way, resulting in 4x4 =16 blocks), n=2 (in total the sample consists of 32 stacks of 15 or 16 (depending on which virtual model was used) increments each).

Hence, a total of 30 x 3 = 90 sampling exercises were performed in the first computer experiment, each of which was repeated 50,000 times.

Because the number of increments in each sample was rather low during the first computer experiment (either 150 (=15x10), 160 (=16x10), 480 (=15x32) or 512 (=16x32) increments per sample), a second computer experiment was performed to study the behavior of the variance estimators as a function of the sample size (the number of increments per sample). Because of the increase in computational time per sample, only three populations (gas01, dpa03 and rok10) were selected for this second computer experiment. Of these models, gas01 is a model with a strong level of large scale heterogeneity, while the model rok10 does not have a high degree of large scale heterogeneity. The sampling method was again stratified simple random sampling.

For gas01 and rok10, the number of blocks was set to 16. The number of vertical stacks included in a sample per block varied: 2, 4, and 8. As a consequence, the number of increments for dpa01 in the second experiment per sample varied from 480 (=15x16x2) to 1,920 (=15x16x8). The number of increments for rok10 varied from 512 (=16x16x2) to 2,048 (=16x16x8).

For dpa03 the number of blocks was set to 1 and the number of vertical stacks per block varied: 10, 20, 40, 80, and 160. As a consequence, the number of increments per sample for dpa03 in the second experiment varied from a minimum of 160 (=16x10) to 2,560 (=16x160)

3.3 Results

The results of the first experiment are presented for all thirty virtual populations. As described in section 3.1, these virtual populations model three actual populations at different levels of spatial segregation. The three actual populations are denoted by "gas", "dpa" and "rok". The virtual populations are denoted by "gas01" to "gas10", "dpa01" to "dpa10", and "rok01" to "rok10" for "gas", "dpa" and "rok" respectively. The number that ranges from 1 to ten indicates the degree of large scale heterogeneity: 1 is the highest degree of large scale heterogeneity and 10 the lowest.

For each scenario (defined by its unique combination of virtual population number, n and N_{block}), a list was obtained of 50,000 independent estimates for c_{sample}, V_{HT}, V_{AD}, $V_{Poisson}$ and V_{SYG}. Using (2) resulted in a numerical value for $V_{est}(c_{sample})$. In view of the high number of replications (N=50,000), it is expected that $V_{est}(c_{sample})$ is very close in value to $V(c_{sample})$.

The Mean Squared Error (MSE) of each estimator was calculated (using the 50,000 values for each estimator in each scenario) as the mean of the squared difference of V_X (where X can be HT, SYG, AD1 or Poisson) and the “true” variance $V(c_{sample})$ (which was approximated by $V_{est}(c_{sample})$).

For both experiments, it was observed that for each given sample the three estimators V_{SYG}, V_{HT} and V_{AD} resulted in three numerically exactly identical values, but $V_{Poisson}$ resulted in a different value (and the difference can vary from sample-to-sample). Using the values of C_{ij} given in section 3.2 it was then also mathematically proven that $V_{SYG} = V_{HT} = V_{AD}$ for the sampling design used here (the detailed proof is not given here). The equality $V_{SYG} = V_{HT} = V_{AD}$ follows from the constant number of increments per block, the sampling of all block, and the specific values of C_{ij} used here. For non constant number of increments per block, not sampling all blocks, or different C_{ij} values, V_{SYG}, V_{HT} and V_{AD} do not have to be equal.

3.3.1 Results for the First Computer Experiment

High Mean Squared Errors were observed during the first computer experiment. The MSE for V_{HT} (and also of V_{AD1} and V_{SYG}) varied from 37% to 2973%. The MSE of $V_{Poisson}$ varied from 33% to 2956%. An overview of all results is presented in Figure 3.

The results indicate that $V_{Poisson}$ structurally has a slightly lower MSE than V_{HT}, for the conditions of the first computer experiment. However, when the MSE of a variance estimator is larger than 100% the practical usefulness of the variance estimate has most likely disappeared. Therefore, the result (that the MSE for $V_{Poisson}$ is lower than for V_{HT} under the conditions of the first computer experiment) should not be used to falsely conclude that $V_{Poisson}$ should become the preferred estimator in practice. The MSE of an estimator may also depend on the sample size (where generally lower MSE’s are expected at higher sample sizes).

One possible attempt to deal with the problem of high MSE’s associated with a variance estimator would be to select a larger sample size (i.e. more increments), because selecting a larger sample size can lead to a lower MSE of a variance estimator. This is studied in the second computer experiment. It is noted that selecting larger sample sizes also serves a more immediate interest: the MSE of c_{sample} can be decreased by increasing the sample size, because $V(c_{sample})$ generally decreases with increasing sample size. It is also noted that the MSE of V_{HT} or $V_{Poisson}$ may be high, while at the same time the MSE of c_{sample} may be low. Vice versa is also possible: low MSE of V_{HT} or $V_{Poisson}$, while at the same time a high MSE of c_{sample}.

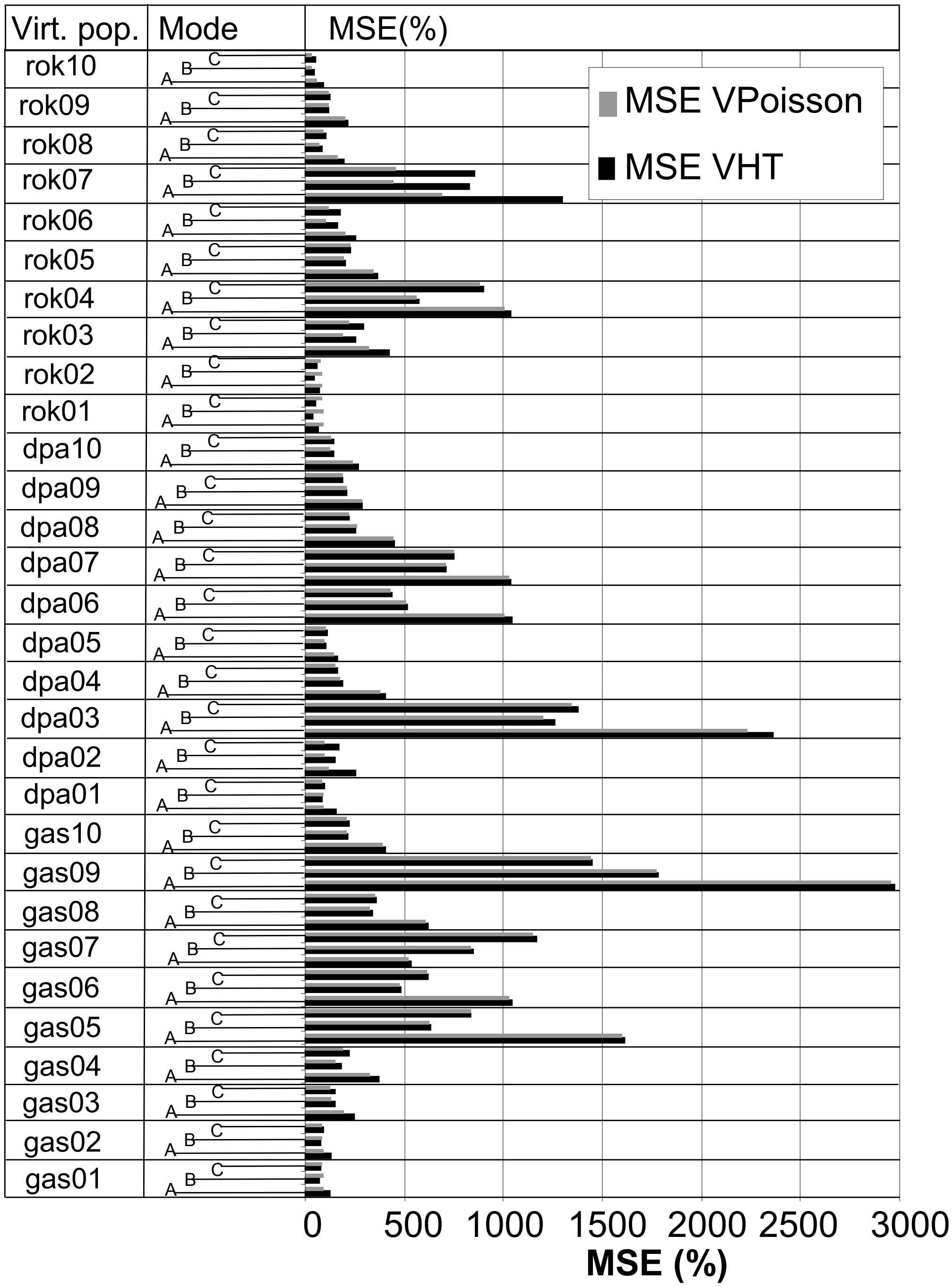


Figure 3. The Mean Squared Errors (MSE's) of V_{HT} and $V_{Poisson}$ expressed as a percentage of the sample-to-sample variance of c_{sample} for the 90 scenarios studied during the first computer experiment. The sampling modes are: A = (N_{Block}=1, n=10), B=(N_{Block}=4, n=8), B=(N_{Block}=16, n=2).

3.3.2 Results for the Second Computer Experiment

It was observed that the MSE (when expressed as a percentage of $V(c_{sample})$) decreases when the sample size increases for V_{HT}, V_{AD} and V_{SYG}, while for $V_{Poisson}$ the MSE remained approximately constant in one of the three scenarios studied during the second computer experiment. This implies that the MSE's of V_{HT}, V_{AD} and V_{SYG} can be made arbitrarily small by selecting a large enough sample. Despite the high MSE's observed in the first computer experiment, the second experiment therefore demonstrates the potential practical usefulness of V_{HT}, V_{AD} and V_{SYG}. Based on the results, it is recommended that $V_{Poisson}$ is not used in practice. The results are graphically depicted in Figure 4, 5 and 6.

Mathematically, the results can be explained as follows: because $V_{Poisson}$ neglects the influence of C_{ij} on the variance it is likely to be biased in most circumstances, while V_{HT} is unbiased (for the scenarios described in this study). The MSE is composed of a bias component and a variance component. While it can be expected that the variance component reduces with increasing sample size, the bias component does not have to decrease with increasing sample size. This is a possible explanation for the observed trends in Figure 4, 5 and 6.

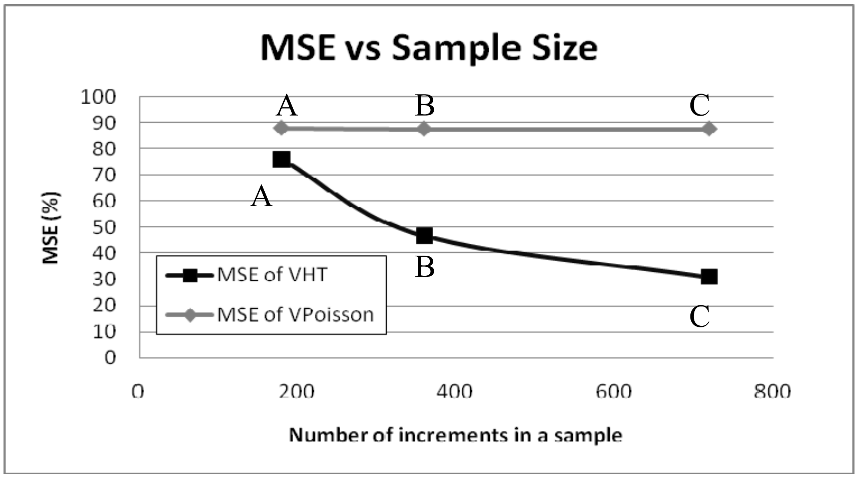


Figure 4. The Mean Squared Errors (MSE's) of V_{HT} and $V_{Poisson}$ expressed as a percentage of the sample-to-sample variance of c_{sample} for the scenarios studied during the second computer experiment. Virtual population = gas01. The sampling modes are: A = (N_{Block}=16, n=2), B=(N_{Block}=16, n=4), C=(N_{Block}=16, n=8).

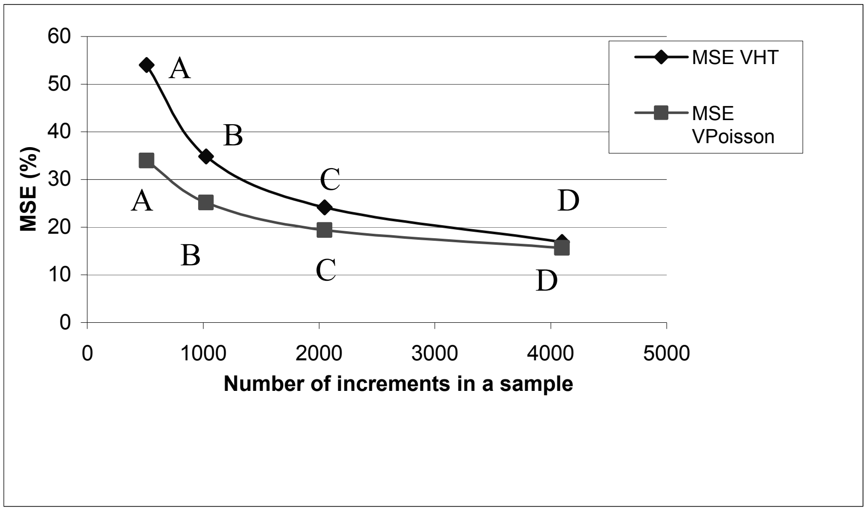


Figure 5. The Mean Squared Errors (MSE's) of V_{HT} and $V_{Poisson}$ expressed as a percentage of the sample-to-sample variance of c_{sample} for the scenarios studied during the second computer experiment. Virtual population = rok10. The sampling modes are: A = (N_{Block}=16, n=2), B=(N_{Block}=16, n=4), C=(N_{Block}=16, n=8), and D=(N_{Block}=16, n=16).

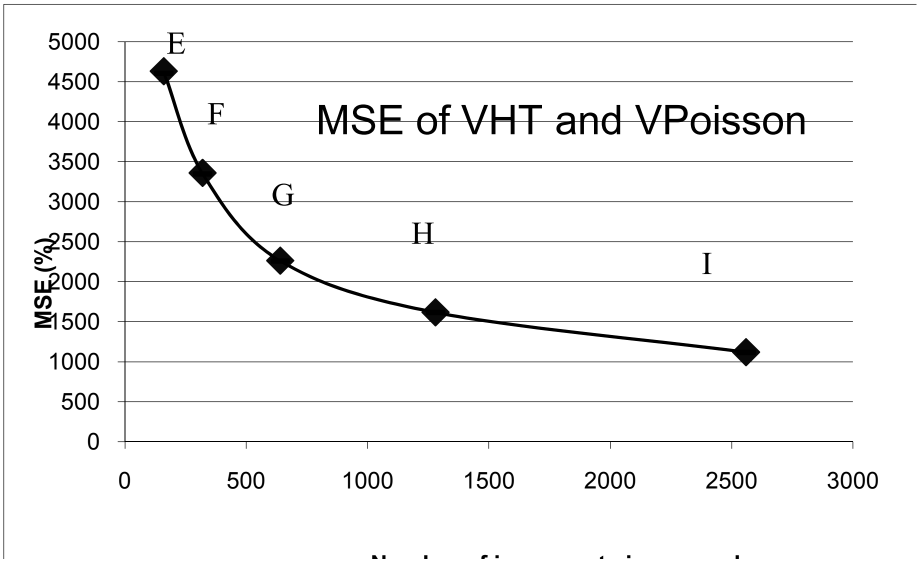
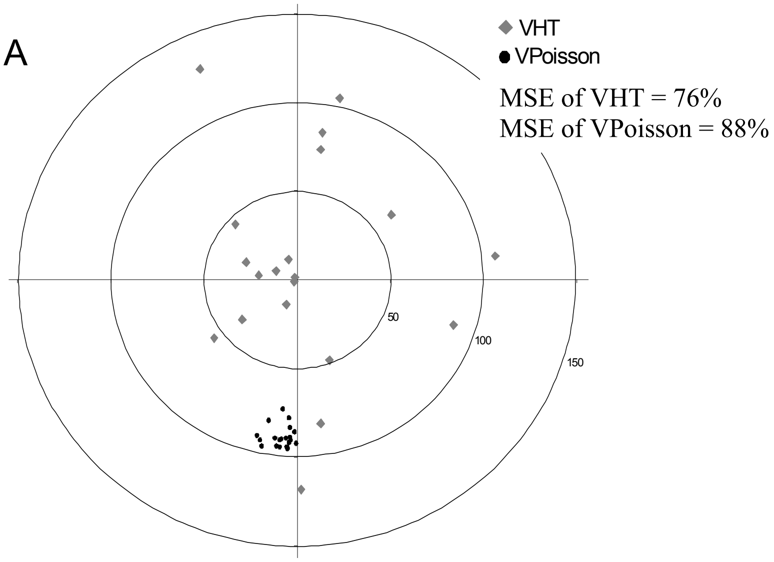


Figure 6. The Mean Squared Errors (MSE's) of V_{HT} and $V_{Poisson}$ expressed as a percentage of the sample-to-sample variance of c_{sample} for the scenarios studied during the second computer experiment (V_{HT} and $V_{Poisson}$ resulted in almost identical values of MSE). Virtual population = dpa03. The sampling modes are: E = (N_{Block}=1, n=10), F=(N_{Block}=1, n=20), G=(N_{Block}=1, n=40), H=(N_{Block}=1, n=80) and I=(N_{Block}=1, n=160).

In Figure 7, the error of estimation of V_{HT} and $V_{Poisson}$ (expressed as a percentage of $V(c_{sample})$) of 20 samples are graphically depicted on a "dart board". The samples where taken from "gas01" using the same sampling modes A, B and C as in Figure 4. Figure 7 clearly illustrates the presence of bias in $V_{Poisson}$ and the reduction of the error (and hence the MSE) of V_{HT} with increasing number of increments per sample.



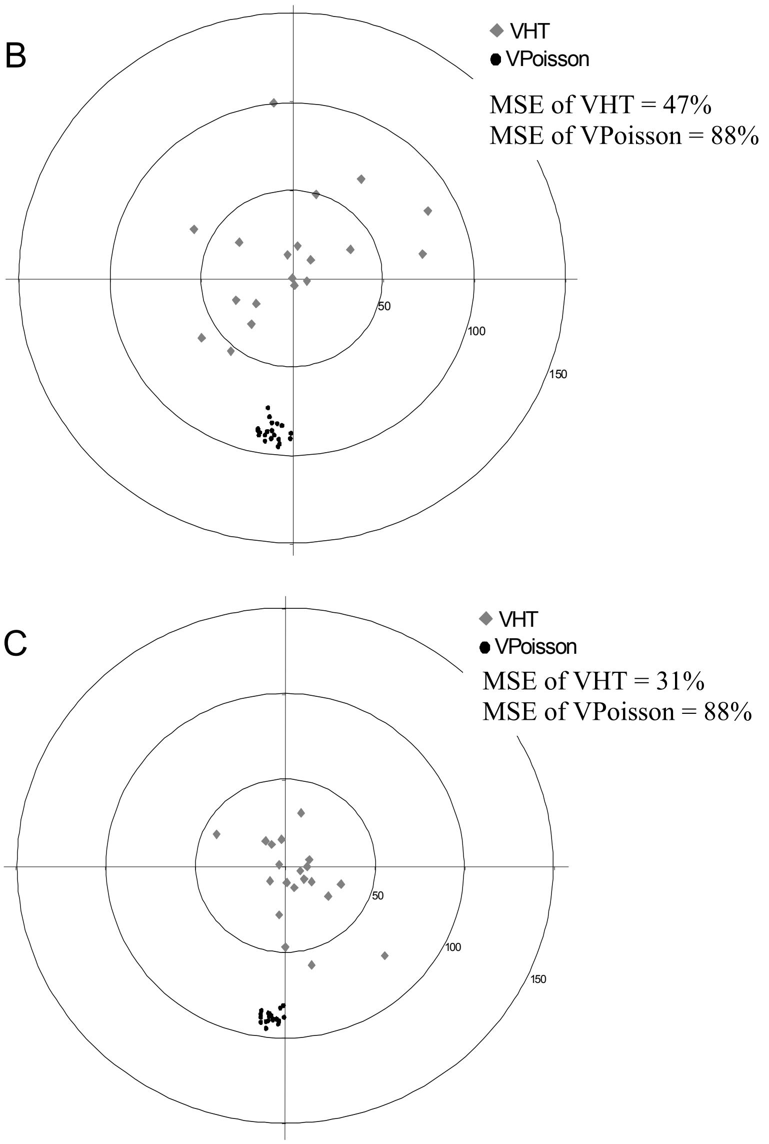


Figure 7. The errors (expressed as a percentage of V(c_{sample})) of V_{HT} and $V_{Poisson}$ for the same virtual population and sampling modes A, B and C as in Figure 4. The distance of a point to the origin represents the magnitude of the error. Hence the analogy with a dart board: a point at the origin means that no error is made when the corresponding sample and estimator are used to estimate the variance. Three circles are depicted for each "dart board": a circle indicating the region with error=50%, 100% and 150%. As the number of increments per sample increase going from A to C, the scatter of V_{HT} becomes more centered in a region close to the origin.

4. DISCUSSION

This study focused on the variance estimators and their MSE. The primary question of whether or not a particular sampling mode will yield samples of sufficiently low sampling variance was not directly addressed in this article. But the results of this article are highly relevant for this question, if it leads to the construction of methods for accurate variance estimation.

5. CONCLUSIONS

Virtual populations were used to study the MSE of four variance estimators.
In a first computer experiment (at low sample size) the following results with respect to the MSE (expressed as a percentage of V(c_{sample})) were obtained:

- The MSE of V_{HT}, V_{SYG}, and V_{AD1} ranges from 37% to 2973%
- The MSE of $V_{Poisson}$ ranges from 33% to 2956%

Although these results suggest that $V_{Poisson}$ performs slightly, but structurally, better, the MSE's of all estimators were high. For V_{HT} V_{SYG} and V_{AD1} this was caused by the low sample sizes during the first computer experiment. It is noted that the structurally lower MSE's of $V_{Poisson}$ during the first experiment do not necessarily indicate that $V_{Poisson}$ itself is structurally lower as well.

A second computer experiments demonstrates that the MSE (expressed as a percentage of V(c_{sample})) decreased with increasing sample size for V_{HT}, V_{AD1} and V_{SYG}. This indicates that these estimators are of interest for future practical application (as opposed to $V_{Poisson}$). For these estimators (V_{HT}, V_{AD1} and V_{SYG}) the accuracy of the estimate can be improved by increasing the sample size. For $V_{Poisson}$, the MSE does not always decrease as a function of increasing sample size, because of potential bias in $V_{Poisson}$.

ACKNOWLEDGEMENTS

This work was performed as part of a project supported by the Netherlands Technology Foundation STW, under STW grant 7457. NFI, Deltares, Nutreco, Hosokawa and Organon are members of the users' committee of this project.

REFERENCES

Geelhoed, B. *Sampling of particulate materials - New theoretical approach*, Delft University Press, 200 pp., Delft, 2004.

Lamé F.P.J., T. Honders, G. Derksen, M. Gadella, Validated sampling strategy for assessing contaminants in soil stockpiles, *Environmental Pollution*, 134(1), 5-11, 2005.

Särndal, C.E., B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, 694 pp., New York, 1992.

Sen, A., On the estimate of the variance in sampling with varying probabilities, *Journal of Indian Society for Agricultural Statistics,* 5, 119–127, 1953.

Yates, F., and P. Grundy, Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society B,* 15, 235-261, 1953.

Appendix A

During simple random sampling of vertical stack of particles, every combination of n vertical stacks has an equal probability of being selected. Denoting the number of vertical stacks from which the sample will be taken by N_{st} The probability of becoming part of the sample of each particle is therefore given by n/N_{st}, i.e:

$$\pi_i = n/N_{pop}$$

On the condition that the first particle has become part of the sample, the probability that the second particle (belonging to a different vertical stack of particles) becomes part of the sample is given by (n–1)/(N_{st} –1). The joint inclusion probability of the pair is therefore n(n–1)/(N_{st}

$\times(N_{st} -1))=(n/ N_{st})^2(1-(1-q)/(n-q))$, where q is the sampling fraction defined (here) by $q=n/ N_{st}$. Expressed as an equation:

$$\pi_{ij} = \pi_i \ \pi_i \ (1-(1-q)/(n-q))$$

From the latter equality, combined with the definition of C_{ij}, it follows that (for $i\neq j$):

$$C_{ij} = (1-q) / (n-q).$$