Extensions to the Theory of Sampling 2. The Sampling Uncertainty (SU), and SU as alternative to variographic analysis

Bo Svensmark

Department of Plant and Environmental Sciences, University of Copenhagen, Faculty of Science, Thorvaldsensvej 40, DK-1871, Frederiksberg C, Denmark

E-mail: svensmark@plen.ku.dk

SUPPLEMENTARY MATERIALS

Contents

S1 Problems with the segregation parameter *Z*

The problem with the Grouping and Segregation Error: The effect of the grouping and segregation is not only dependent on the distributional heterogeneity of the lot, but more important, on the position of the increments in the sample relative to the spatial distribution of the constituents of the lot.

Fig. S1 gives an example with 3 classes of fragments, a background matrix with very small particles (the white background), one analyte with small particles (small dots) and another analyte with larger particles (big dots). The degree of distributional heterogeneity is identical in the two cases A and B. Assume furthermore that the sample mass is half the mass of the lot, i.e. the sample is the sum either the even or the uneven increments.

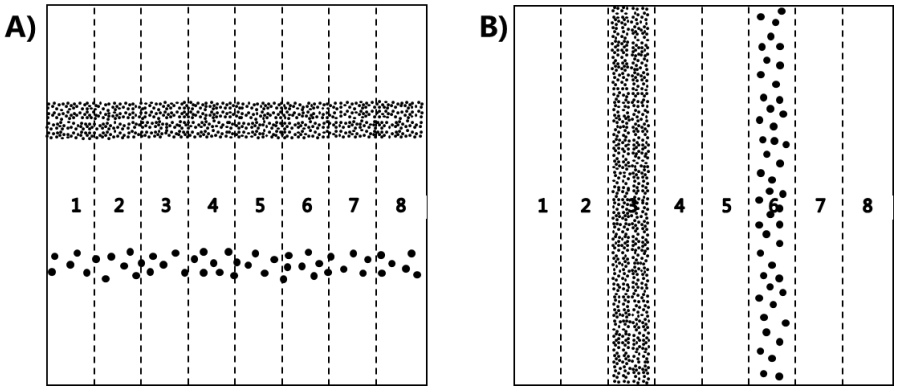

Figure S1. Example with systematic sampling from a lot with A) horizontal and B) vertical stratification of 2 analytes.

In TOS the correct sampling variance (CSE) is expressed by Eq. (1):

$$s^2\left(CSE\right) = s^2\left(SFE\right) + s^2\left(GSE\right) = \left(\frac{1}{M_S} - \frac{1}{M_L}\right)\cdot\left(1+ZY\right)\cdot HI_L \qquad (1)$$

where *Y* is the grouping parameter, which is the average number of fragments in an increment, and *Z* is the parameter for the distributional heterogeneity. In the example, Y and Z would be the same for case A and B and for the small and the big particles because they occupy the same fraction of the lot and have the same distribution pattern.

However, the empirical GSE would be very different in the two cases: in A the *effect* of grouping and segregation on the sampling variance will be very small – only caused by the fluctuations in the level of the borderlines between pure matrix and analyte particles. Conversely, the *effect* of grouping and segregation will be extremely high in B: if the border for the strata can vary randomly relative to the increments, GSE gives a standard deviation of 58 % for both type of particles.

Now if one assumes that the heterogeneity invariant is 0.01 g for the smaller particles and 5 g for the larger ones, *Z* needs to be ~5/0.01 = ~500 times higher for the small particles than for the larger ones. So even if the distribution due to grouping and segregation in the lot has exactly the same magnitude for small and big particles, and for case A and B, the effect on GSE is difficult or impossible to estimate using Eq. (1), because the parameter *Z* depends on FSE in an unpredictable way. An example from literature is given as below.

An example from literature: A good example of the problems with the coupling of GSE to FSE by use of the factor (1+YZ) to obtain the sampling variance from FSE is the work of Geelhoed et al. [9]. For a model experiment with a riffle splitter FSE was predicted to be $0.96\cdot10^{-7}$, but the observed variance was $5.67\cdot10^{-6}$ thus the correction for grouping and segregation was found to be a factor of ZY = ~60. Geelhoed et al. correctly state that the theory of Gy was unable to predict the magnitude of this correction, so instead they used a new model for calculating the observed variance. This model uses the variance and covariance of the number of particles in a series of samples in order to predict the observed variance. However, these variances and the covariance are estimated from the results of the experiments, so this is no more prediction than calculating the correction factor in Gy's formulation. So neither Gy's nor Geelhoed's theory is able to predict the sampling variance – they both use the results of the experiments for the estimation. The Sampling Uncertainty SU cannot be used to *predict* the grouping

and segregation variance in this case, but it can be used to *estimate* it from the degree of uniform mixing of the lot.

The reader may find that a grouping and segregation error of 60 times the fundamental sampling error is far too high to be realistic. But a factor of 60 for GSE vs FSE is not unrealistic at all. In the Geelhoed paper, the reason is that the correct FSE is extremely small ~10^{-7} corresponding to a relative standard deviation of only 0.03 %. The observed standard deviation was 0.48 % meaning that virtually all variation is due to grouping and segregation and not fundamental sampling error. The grouping and segregation error for a riffle splitter is caused by the distribution of the material between the chutes, i.e. in the longitudinal direction of the input tray. In order to get a sampling standard deviation of 0.48 % the standard deviation for the number of particles in a single chute of the Riffle splitter with 18 chutes must be ~1.5 %, which is a very likely order of magnitude for a well mixed material of two size spherical particles. If the grouping and segregation error should be equal to FSE, this standard deviation needs to be as low as ~0.15 % and for GSE to be negligible, it must be <0.05 %. Now it may be possible to get a uniform distribution of particles with a standard deviation between the chutes of 1.5 % but hardly 0.15 % or 0.05 %, so the experiment is deemed to end up with a grouping and segregation error of >> FSE. To use the traditional approach (GSE = (1+YZ)·FSE) we need to estimate a YZ of ~60. To use Geelhoeds method, the variance and covariance of the results must be estimated. To use the SU the result can be estimated using a standard deviation for the difference in concentration between chutes, which is more accessible for qualified estimates.

S2 Mathematical sampling boundary

The benefits of the circular mathematical sampling operation are explained in Fig. S2 for a case similar to that in Fig. S1, i.e. the degree of segregation or grouping is exactly the same, only the position is different:

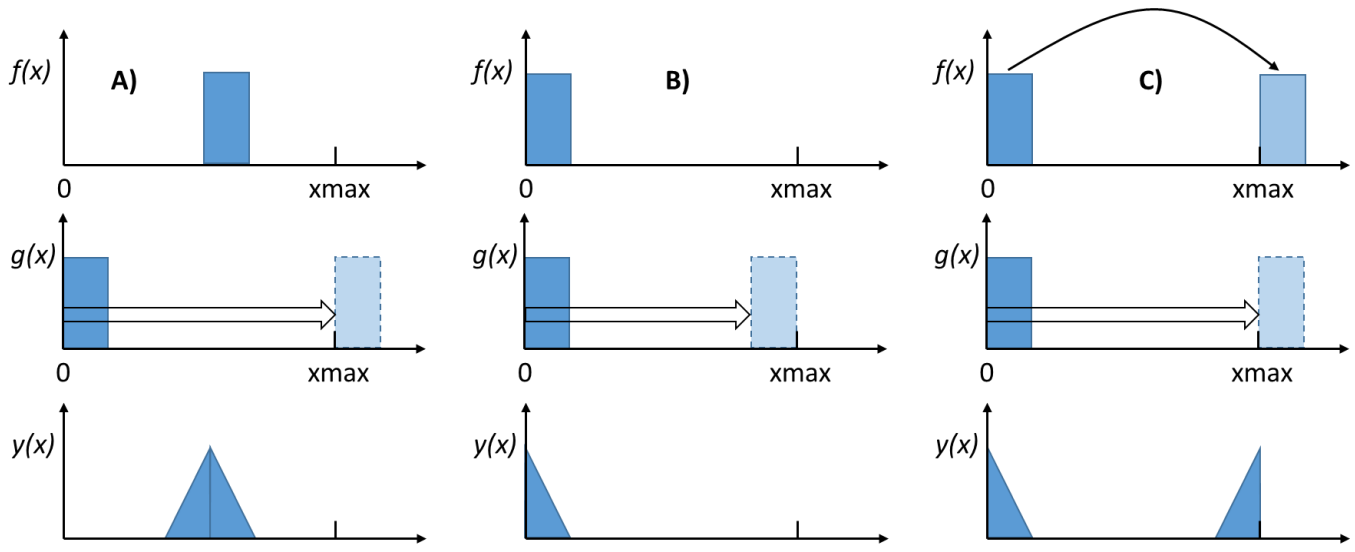

Fig. S2. Sampling of an analyte only present in a part of the lot *f(x).*The sampling function is *g(x)* and *y(x)* is the concentrations in the samples where *x* indicates the left border of the increment. A) Analyte

present inside the lot. B) Analyte present at the start of the lot, and sampling operation only within the lot. C) Analyte present at the start of the lot and sampling operation starting at all positions within the lot. The concentrations in C) for $x > xmax$ is a copy of the start of $f(x)$ as indicated by the arrow.

As can be seen the average concentration in the samples $y(x)$ *in* A) and C) are identical and double that obtained in B). When the standard deviation of sampling is calculated from the concentrations, they will be correct in A) and C) but wrong in B). The proposed mathematical sampling operation is identical to circular convolution integration $y(x) = f(x) \cdot g(x)$, so the concentration in the potential samples could be obtained by circular convolution of the concentrations in the lot with a sampling function $g(x)$ describing the increments. For stratified random sampling and random sampling the sampling functions $g(x)$ are not constant, but vary randomly. For this reason convolution is only a practical approach for single increment sampling and random systematic sampling, and this approach is not used in the Excel sheets.

S3 Variograms and SU

Variograms: A variogram is the semivariance, $v(j)$, as function of the lag, j, i.e. the distance between samples:

$$v(j) = \frac{1}{2 \cdot (N - j)} \sum_{i=1}^{N-j} \left(x_i - x_{i+j} \right)^2 \qquad (1)$$

$v(j)$: Semivariance
j: Lag
N: No of samples
x: Concentration or heterogeneity, h see Eq. (2 in the paper)

The sampling uncertainty is estimated by integration of the variograms as described in [1, 2, 9], with a small correction: The predicted uncertainty for lag $j = 1$ is identical to the nugget effect for systematic sampling, which may be wrong as shown by Heikka and Minkkinen [10]. A simple way of fixing part of this problem is to approximate the variance for systematic sampling for $j = 1$ by a linear extrapolation from the variance $j = 2$ and 3, i.e. $s_1^2 = 2s_2^2 - s_3^2$. This fixes the problem with a discontinuity between $j = 1$ and $j > 1$, but not the effects of small nugget effects in general.

Variograms and SU: The weights for data in comparison of variograms and SU can be seen in Fig. S3.

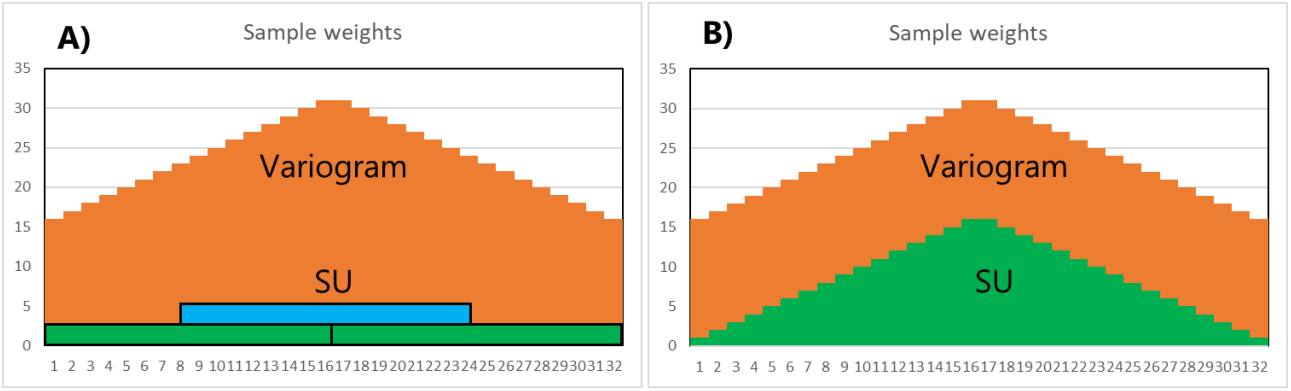

Figure S3. Sample weights for an example with 32 data points. For SU each point is included once for one run as indicated by the blue box in A). When 2 boxes are used to cover the whole range 1 to *xmax* (green

boxes A) the weights will be 1 for all points, and when 3 boxes are used (blue + green boxes A) the weights resemble that of the variogram (double weights at the centre). Finally when all possible runs are used (Fig. S2 B, starting at all positions 1 to *xmax*/2+1) the weights will be even more extremely skewed than for the variogram (1 to *xmax*/2 -1, green profile in B).

Unfortunately, it is not possible to get the same weights for the samples in the variogram and in SU. In the paper the method shown in Fig. S3 B) is used. The weights for variograms increase from *xmax/*2 to *xmax*-1 and for SU from 1 to *xmax/*2-1. This will have an effect when the standard deviation in the data is not approximately uniform along *x*.

S4 Trends in data

For a linear trend without noise, Fig. S4, the mean of the double integral w'(j) is exactly double as the size of the mean of the single integral at the half lag w(j/2), Table S1. SU and variograms give exactly the same uncertainty for stratified random and random sampling, but the uncertainty for random systematic sampling is equal to zero from variograms. For SU systematic sampling has a higher uncertainty compared to stratified random sampling, because stratified random sampling in average would cover a smaller range of the data than strict systematic sampling.

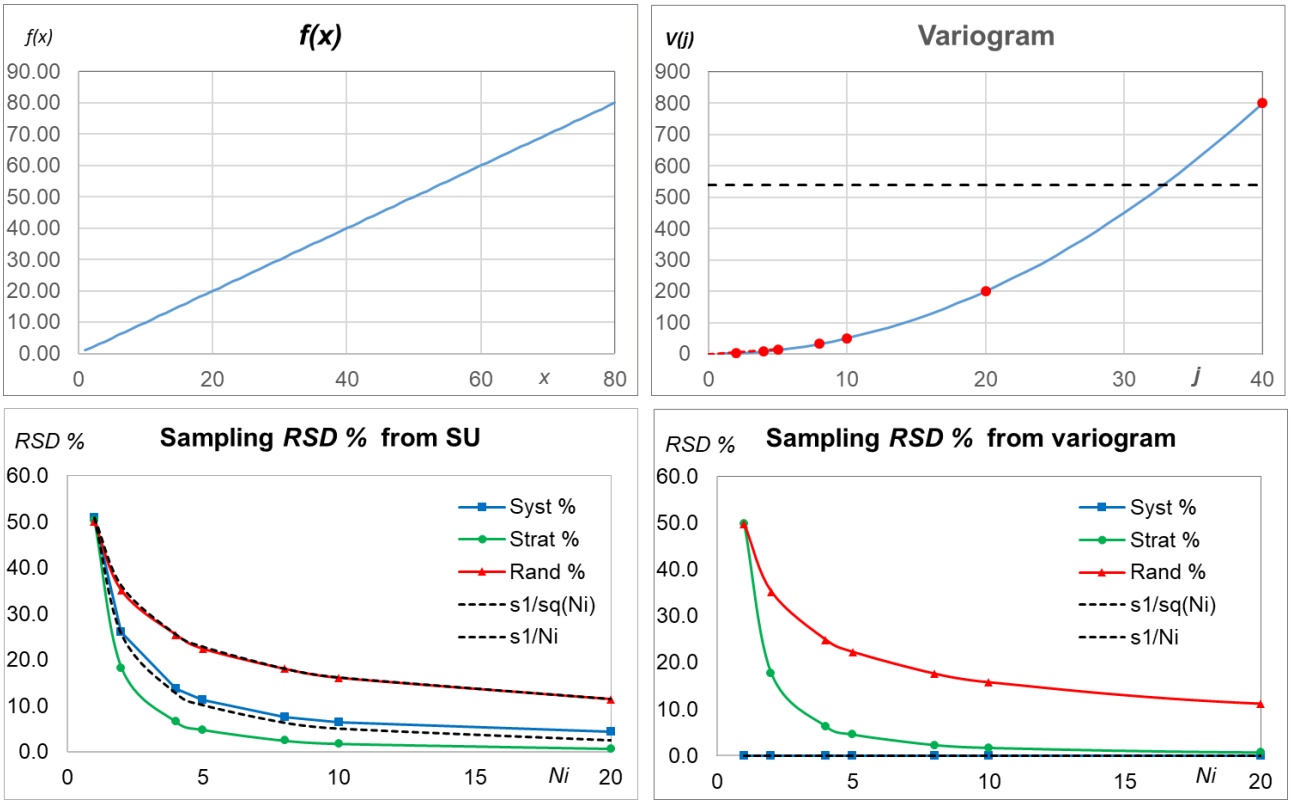

Fig S4. Data, variogram and predictions for a linear trend.

Table S1. Integration of variogram for a linear trend (extracted from workbook SU_Vario_3.0.xlsx) .

j	V(j)	S(j)	w(j)	S'(j)	Var(strat) w'(j)	2w'(j/2)	Var(syst) W(sy)
1	3	2	2	1	2	2	0
2	12	9	5	6	3	3	0
3	27	29	10	25	6	6	0
4	49	67	17	73	9	9	0
5	76	130	26	171	14	14	0
6	110	223	37	348	19	19	0
7	149	352	50	635	26	26	0
8	195	524	66	1073	34	34	0
9	247	745	83	1708	42	42	0
10	305	1021	102	2591	52	52	0
11	369	1358	123	3781	62	62	0
12	439	1762	147	5341	74	74	0
13	515	2239	172	7341	87	87	0
14	597	2795	200	9858	101	101	0
15	686	3437	229	12974	115	115	0
16	780	4170	261	16778	131	131	0
17	881	5001	294	21363	148	148	0
18	988	5935	330	26831	166	166	0
19	1100	6979	367	33288	184	184	0
20	1219	8139	407	40847	204	204	0
21	1344	9421	449	49627	225	225	0
22	1475	10831	492	59753	247	247	0
23	1613	12375	538	71356	270	270	0
24	1756	14059	586	84572	294	294	0
25	1905	15889	636	99547	319	319	0
26	2061	17872	687	116427	344	344	0
27	2222	20014	741	135370	371	371	0
28	2390	22320	797	156537	399	399	0
29	2564	24797	855	180095	428	428	0
30	2743	27450	915	206219	458	458	0
31	2929	30287	977	235087	489	489	0
32	3121	33312	1041	266886	521	521	0
33	3320	36533	1107	301808	554	554	0
34	3524	39954	1175	340052	588	588	0
35	3734	43583	1245	381821	623	623	0
36	3951	47426	1317	427325	659	659	0
37	4173	51488	1392	476782	697	697	0
38	4402	55775	1468	530413	735	735	0
39	4636	60294	1546	588448	774	774	0
40	4877	65051	1626	651120	814	814	0

If random noise is added to the linear trend, a similar behavior is observed, Fig S5, S6 and S7 : SU and variograms give the same prediction for stratified random and random sampling, and the uncertainty depends on the trend. For systematic sampling, variograms give the same uncertainty independent of the slope, which is different from SU. Without a slope, Fig. S5, the predictions for systematic sampling are identical for SU and variograms. Note that the predictions for random systematic sampling from variograms are identical in all 3 cases independent of the trend.

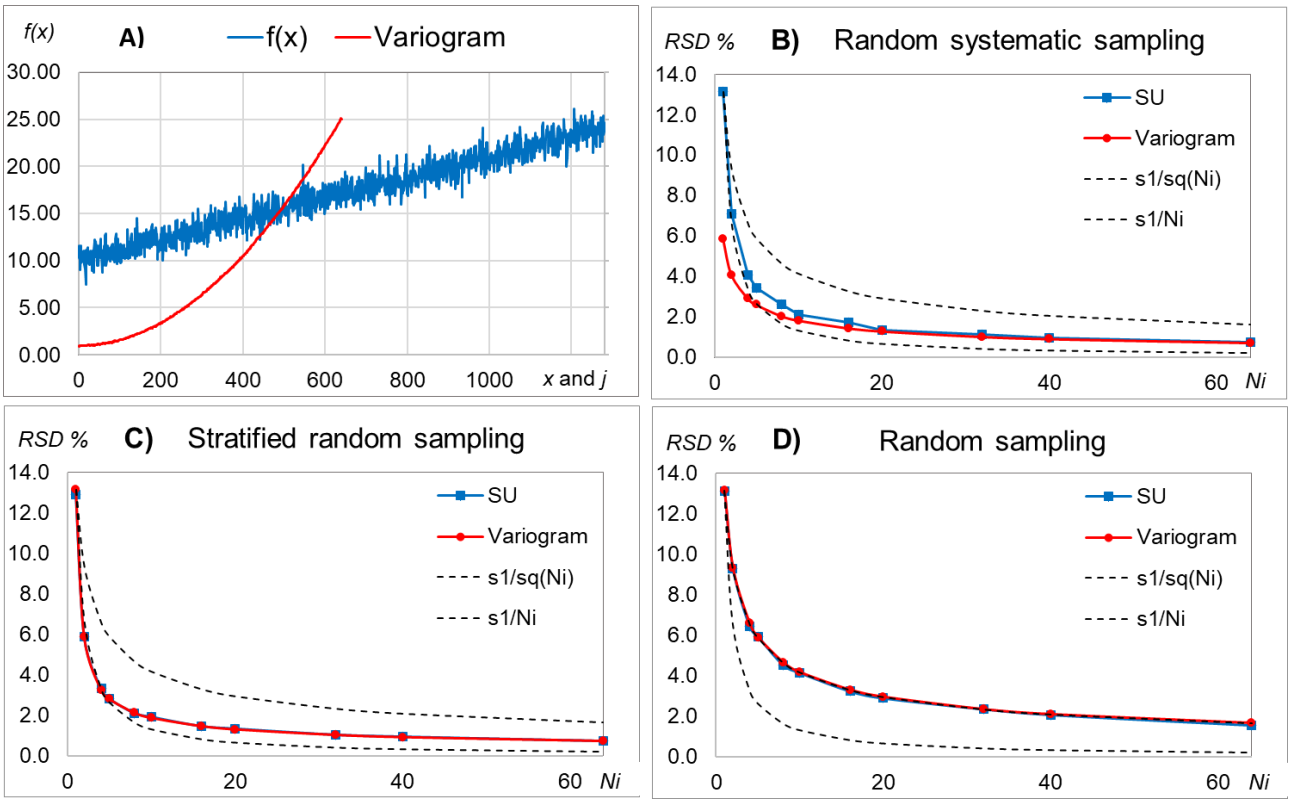

Figure S5: Positive linear trend with noise.

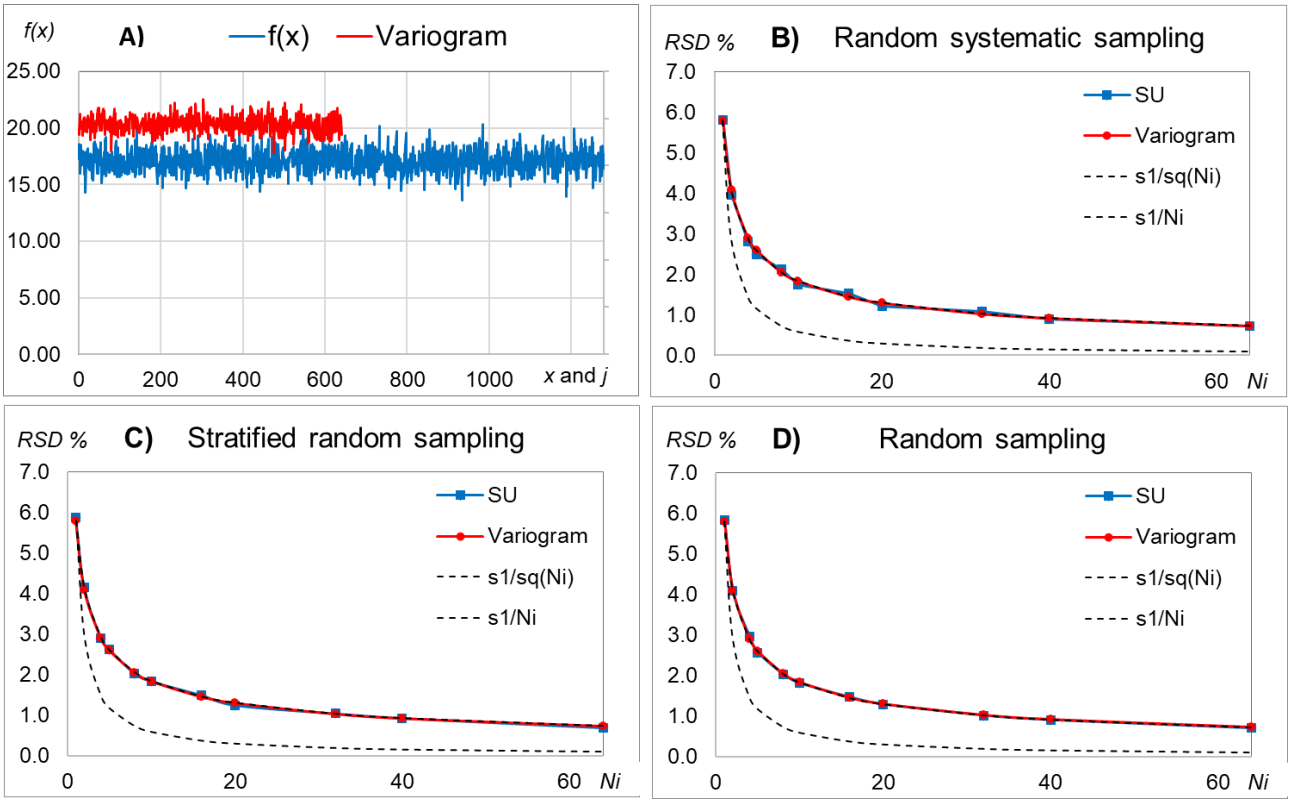

Figure S6: Noise with no trend, slope = 0.

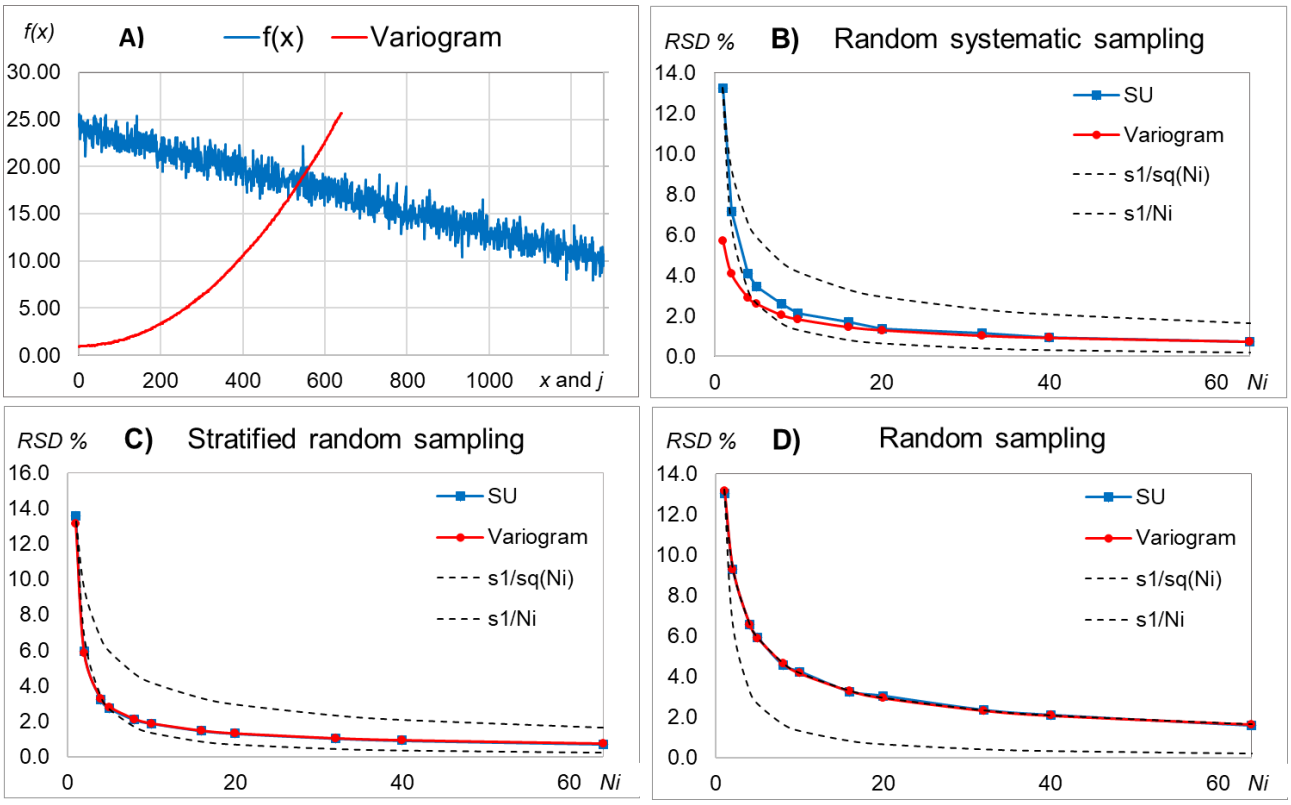

Figure S7. Negative linear trend with noise.

S5 Cyclic variations in variograms

The difference between variograms and SU in a case with a pure cyclic variation with a period of $j = 16$ without added random noise is very high, Fig. S8.

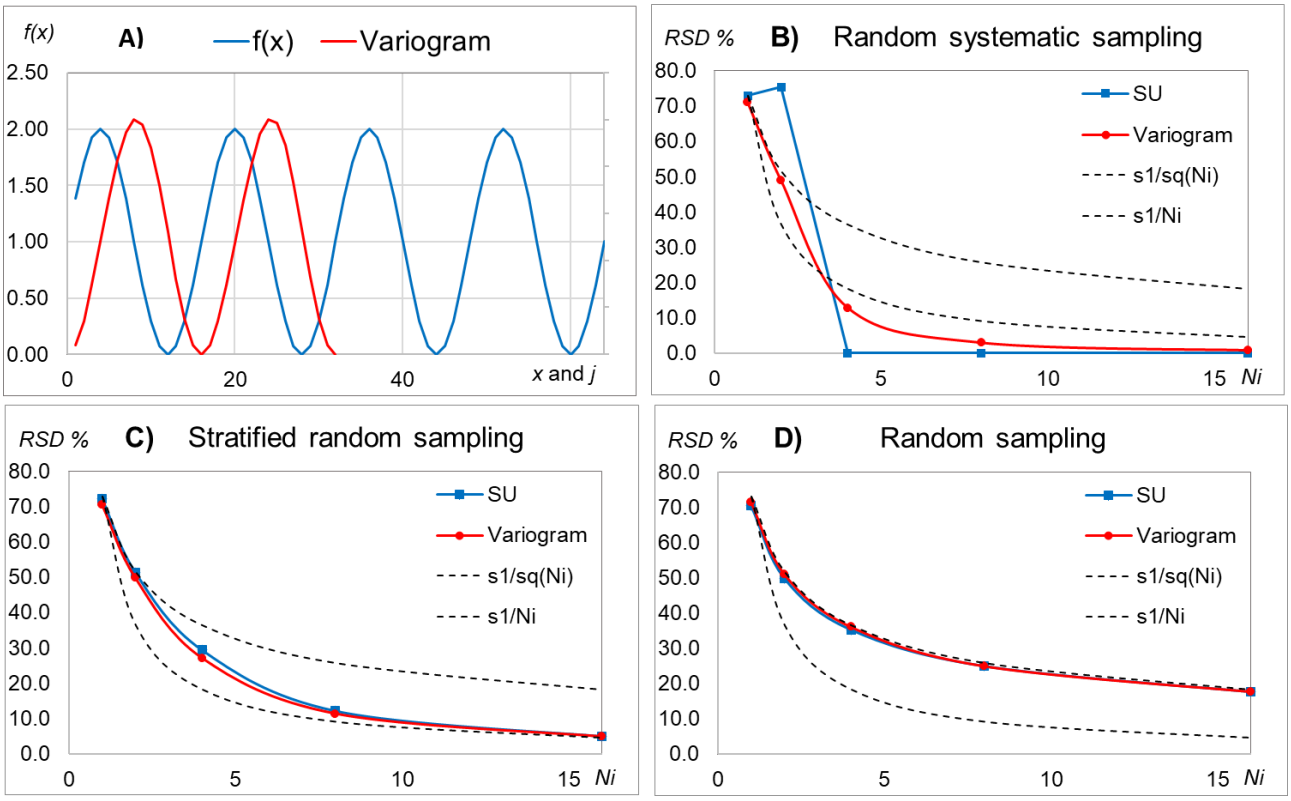

Figure S8. Predicted sampling uncertainty from variogram C) and SU D). A) Raw data and B) variogram.

The SU method correctly predicts that the sampling uncertainty for systematic sampling is at the maximum for sampling periods equal to or longer than the period in the data ($j = 32$ and 16 corresponding to $N = 1$ or 2) and decreases to zero when we sample an even number of increments within each period ($N = 4$, 8 or 16 corresponding to $j = 8$, 4 or 2). In contrast, the results from variogram integration do not reflect the effect of the sampling period relative to cyclic period. For stratified random and random sampling the results are almost identical.

However, the variogram itself can be used to find the good and the bad sampling intervals. The cycles in the variogram always starts at $j = 0$, i.e. when the period is 16 the first minimum will be at $j = 16$ independent of the phase of the cycles in the data, Fig. S9. There may be a small shift due to an increasing semivariance caused by autocorrelation.

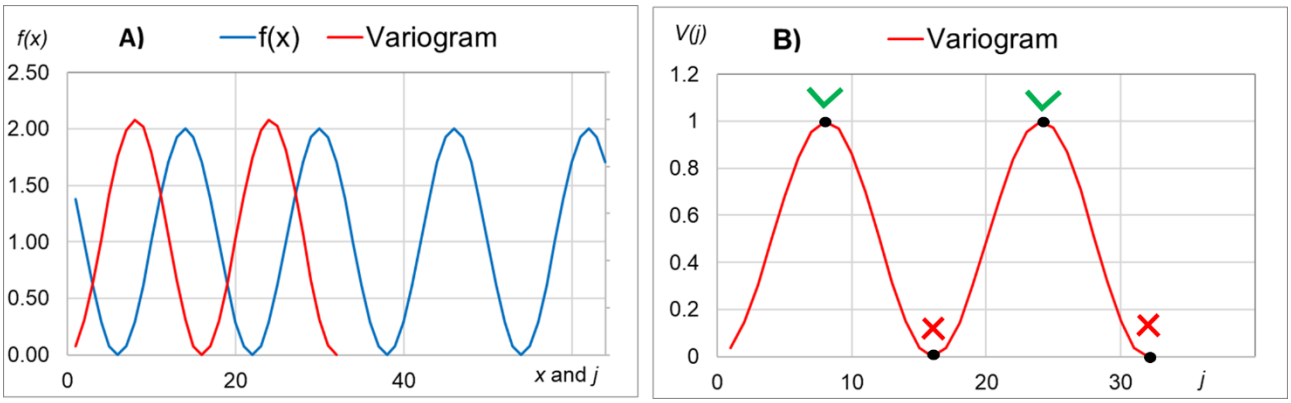

Figure S9. A) Same data as Fig. 5 shifted -6 lags; B) and variogram with indication of good (*j* = 8 and 24) and bad (*j* = 16 and 32) sampling lags.

At the lag with minimum semivariance (*j* = 16 and 32) the variance *between* the increments is at a minimum, but the variance of the mean of the increments, i.e. between the means when sampling is repeated many times, is at a maximum. And *vice versa* at the maximums (*j* = 8 and 24). So the golden rule for random systematic sampling is to sample at the maxima of the cycles in case of cyclic variations. This is true, also for cyclic variations which are not pure sine waves.

If the frequency of the cycles is constant, e.g. caused by daily variations, random systematic sampling is ok, but in the more general case, where the frequency may shift, the only recommended sampling method is stratified random.

S6 Data with low nugget effect and effect of random noise

For comparison of SU to variograms, the variance for a length of *xmax*/2 is calculated for all potential samples in the whole range. For these data with *xmax* = 96 there are 49 different ranges of the half length, giving 49 possible variances, shown in Fig. 10 A) for one increment, lag *j* = 48 and in B) for 24 increments, lag *j* = 2. The dotted lines show the estimate for SU (the average of the variances) and for variogram integration. For 1 increment it is not possible to take any sample from the lot with as low an uncertainty as predicted by the variogram. For 24 increments there are a few ranges, where variograms could be correct, but on average there is a large discrepancy. For *Ni* between 1 and 24 results are intermediate, not shown. Any linear trend in the data was removed before this comparison.

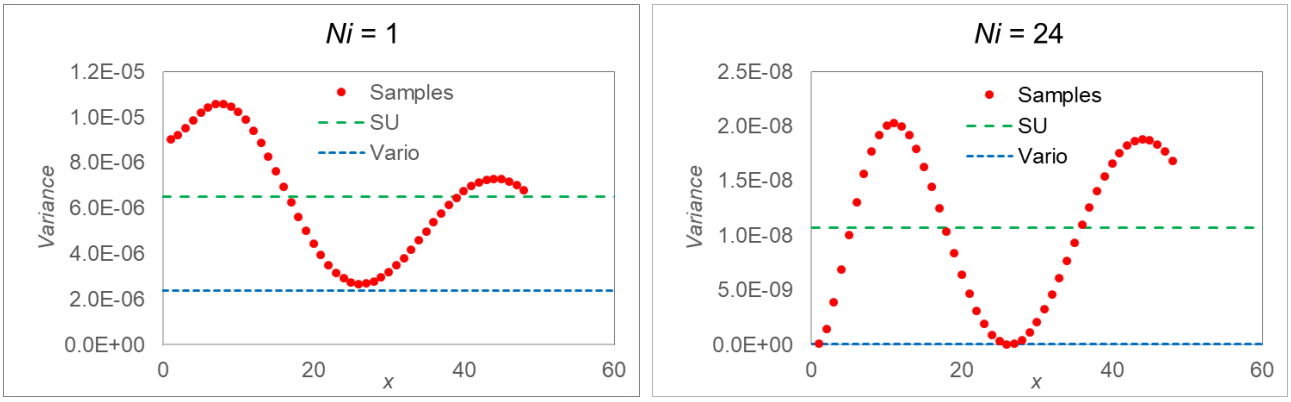

Figure S10. Variances for random systematic sampling predicted by SU (Green line), variograms (Blue line) and the individual mathematical sampled variances for estimation of SU (Red dots).

If random noise is added to the data with low nugget effect in Fig. 5, the nugget effect is > 0 and SU and variogram gives similar results for random systematic sampling, Fig. S11.

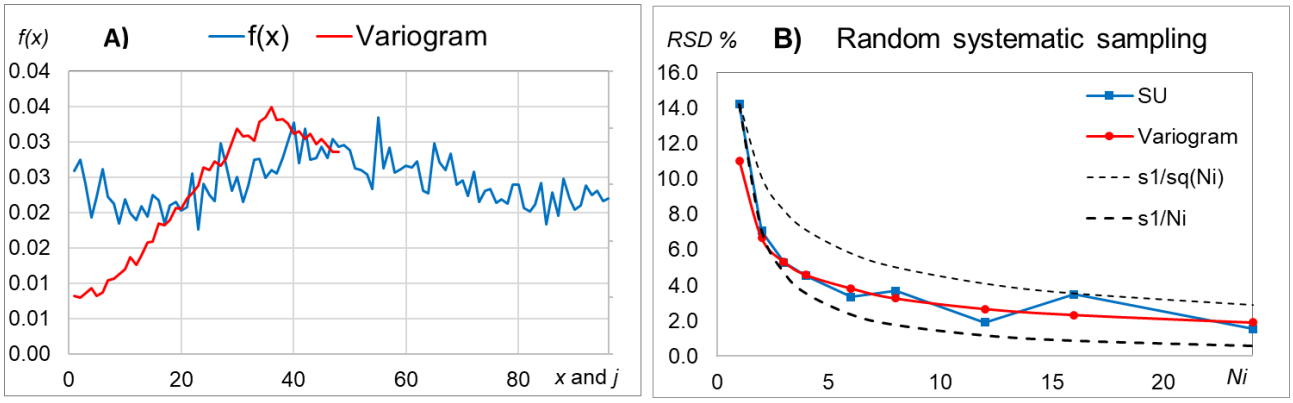

Figure S11. Smooth data with added noise. Predictions of sampling uncertainty (*RSD %)* for variogram integration and for SU estimation. Data from Fig. 5 with added Gaussian distributed noise.

The variogram method is very sensitive to trends in data, and although the trend is not linear, it still affects random systematic sampling for *Ni* = 1. SU on the other hand is very sensitive to cyclic variations, and even a small random cyclic pattern gives periodic deviation from a smooth curve.

S7 SU *vs* variograms for non-stationary standard deviation in data

When the data have separate areas with a high and low standard deviations the predictions for both SU and variograms shifts as a function on the position of these areas. Fig. S12 shows two sets of identical data, the only difference is that the pattern is shifted 30 units.

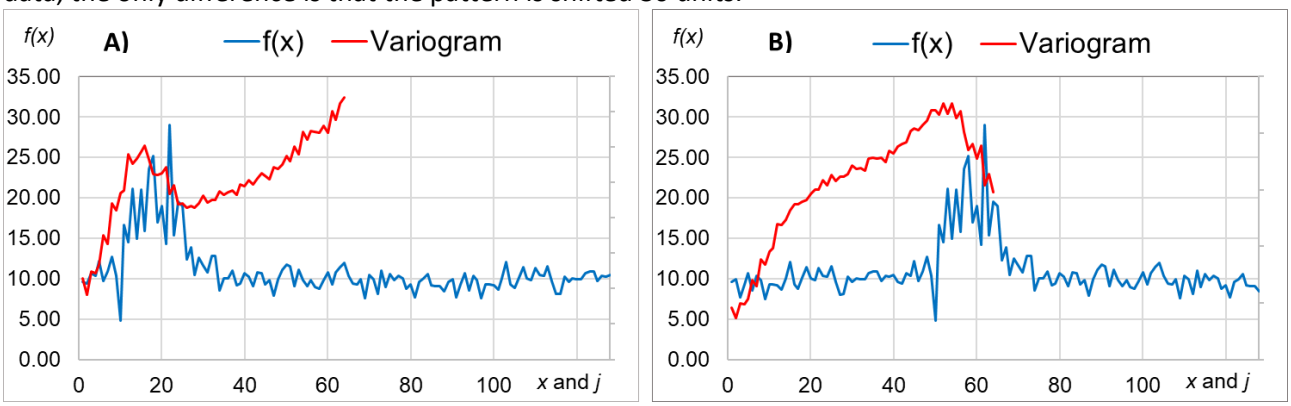

Figure S12. Data and variogram for a data set with variations in standard deviation. The data in case A) and case B) are identical apart from a shift of 30 units.

If the weights of the data for SU and variograms were uniform, the predictions for both data sets should be identical, but they are not, Fig. S13.

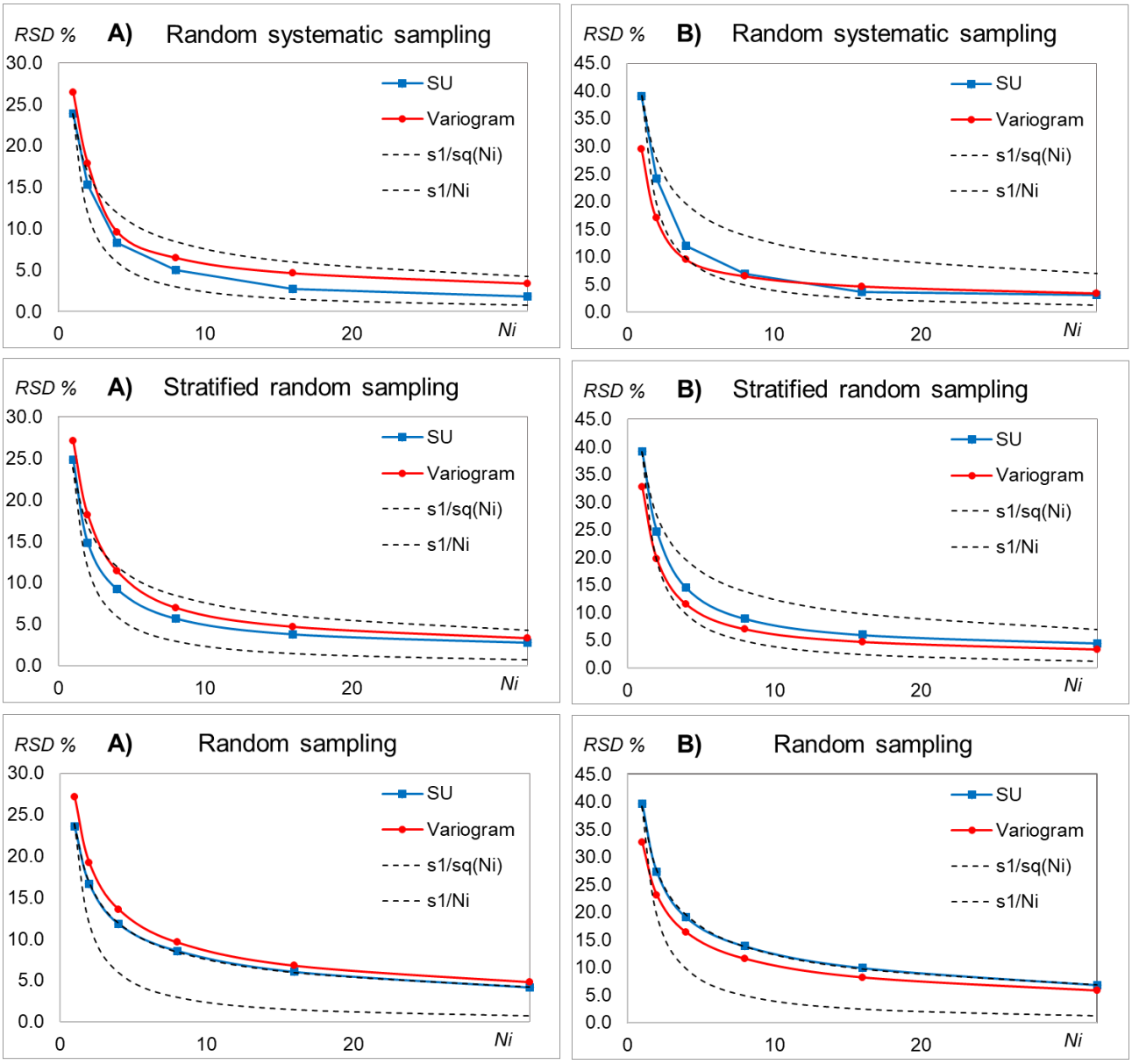

Figure S13. Predicted *RSD %* for data set case A) left and case B) right.

Both methods predict lower sampling uncertainty for case A than for case B. SU and variograms have a too high weight for the data in the centre, but the weights are more skewed for SU than for variograms. For data set A the standard deviation at the centre is lower than the average standard deviation, thus SU underestimates the uncertainty to a higher degree than variograms. For data set B SU overestimates the uncertainty to a higher degree than variograms.

This last case seems to be a problem for the SU method but this is not the case. The only reason for that the skew weights have been used in SU here is to be able to compare to results from variograms. However, there is not this restriction for the SU method in general, as discussed further in the next sections.

SU for 1-dimensional sampling of the whole length of 1-dimensional data: Contrary to variogram integration, SU can easily be used to predict sampling uncertainty for the whole length of the data. Fig. S14 shows the results for sampling from the range *x* = 1 to 128 of the data in either Fig. S13A or Fig. S13B .

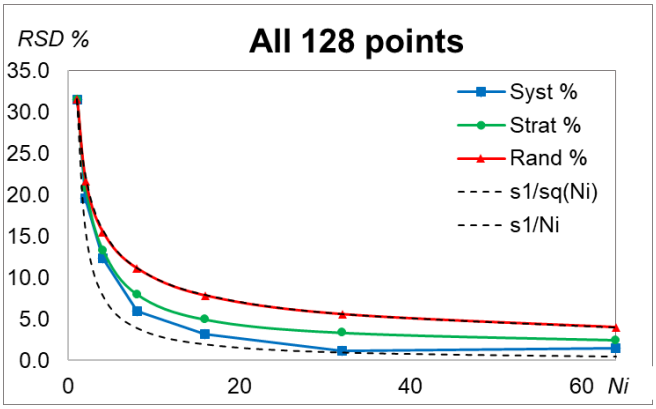

Figure S14. Uncertainty *RSD %* predicted by SU for sampling of the whole range *x* = 1 to 128 of the data in Fig. S13.

The results are identical for the data in Fig. 13A and B (within the uncertainty for the Monte Carlo simulations), thus the results from SU are independent of the non-stationarity as long as data are identical apart from a cyclic shift. This property is very important in the case of cyclic variations, because a shift in the phase of the cycles could otherwise have a huge effect on the results.

The conclusion for the comparison of SU and variograms is that SU gives similar or better results compared to variograms. Also all differences between the SU and variogram results can easily be accounted for by the properties of variogram integration and the different weightings as explained at the beginning of these sections.

In conclusion, SU is a valid method for estimation of sampling uncertainty for 1-dimensional sampling.

S8 SU for 1-dimensional sampling as function of autocorrelation

SU is calculated for different scenarios (Fig.S15 A-F) in order to show the effect of autocorrelation.

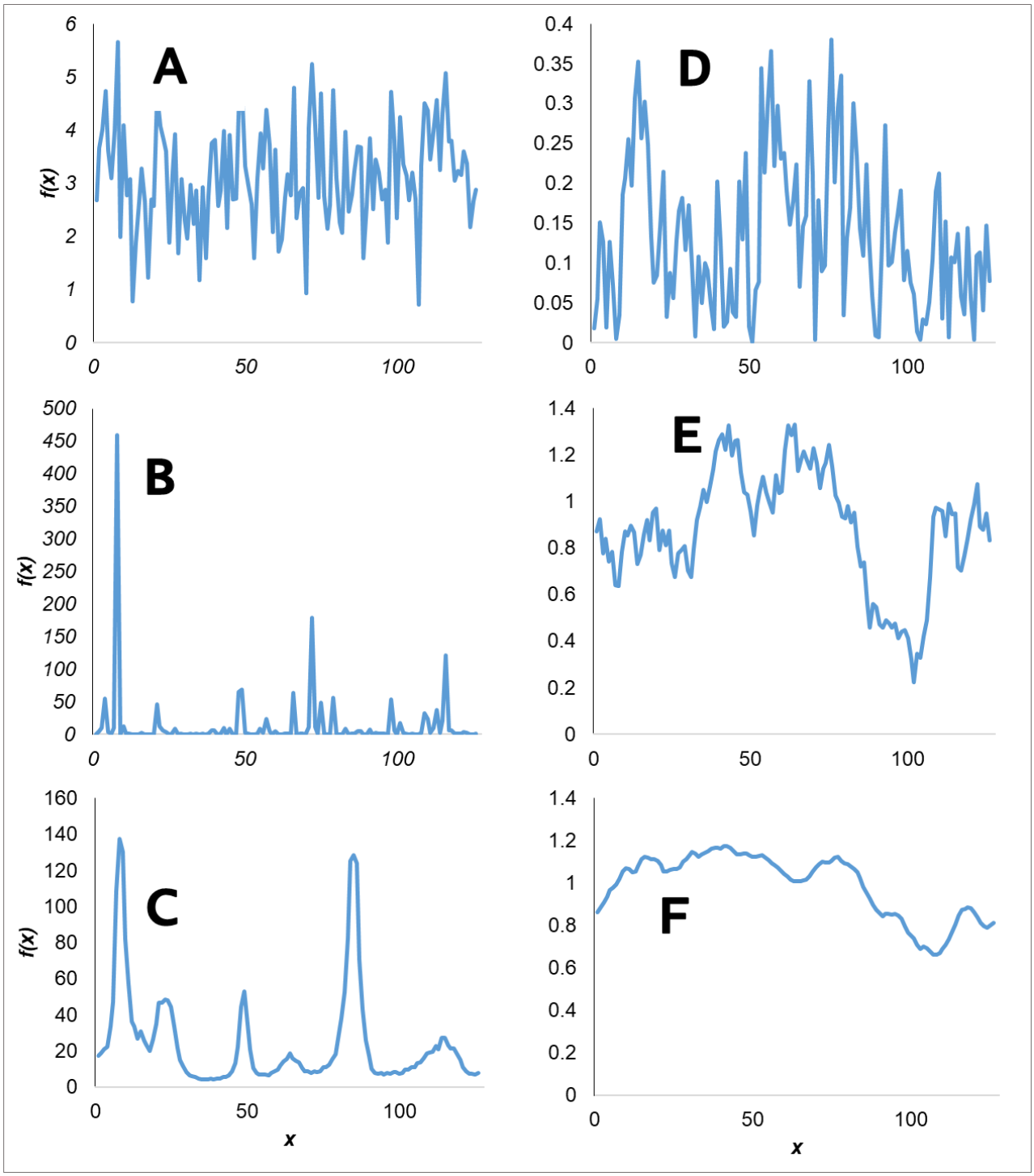

Figure S15. Six scenarios, concentrations *f(x) vs* position *x*, with increasing autocorrelation. A) Random noise with *mean* = 3 and *s* = 1; B) Log-normal random distribution *f(x)* = 10^{A-3}; C) Smooth log-normal random distribution; D) to F): Cases with weak D) to strong autocorrelation F).

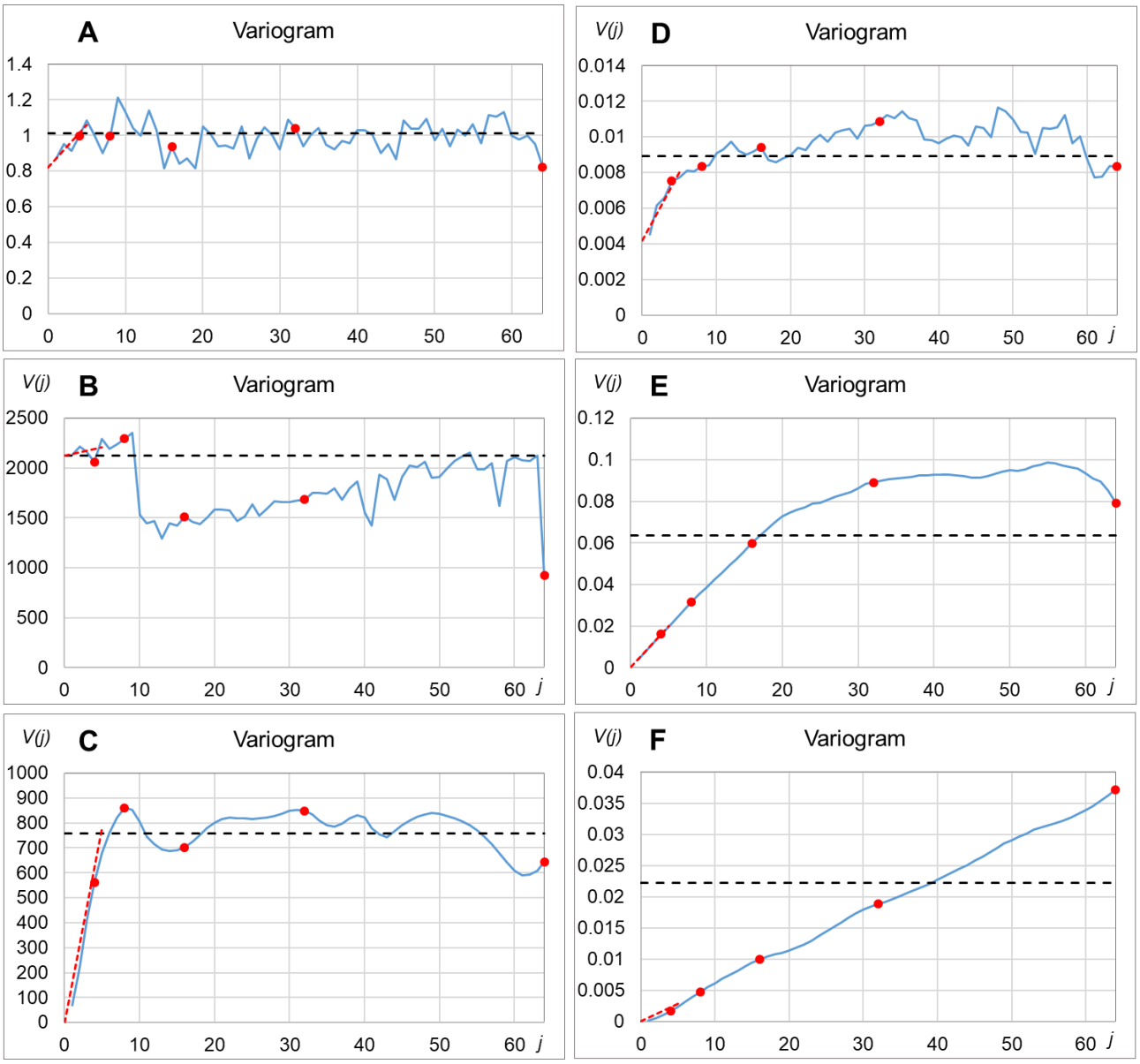

Figure S16. Variograms for the data A-F in Fig. S15

The variograms clearly show that case A and B has no autocorrelation and nugget effect equal to the sill, case C has a strong autocorrelation for a short range, and finally the autocorrelation and the range increases from case D to F. Discrete Fourier Transforms show no cyclic pattern in the data A-F (not shown)

The estimated SU for the scenarios (Fig.S17) show that the number of increments *Ni* and the sampling method (random systematic, stratified random or random) is unimportant for case A and B. When the distribution is random (Gaussian or log-normal) the only factor that affects sampling uncertainty is the sampling ratio (see Fig. S17 and text below).

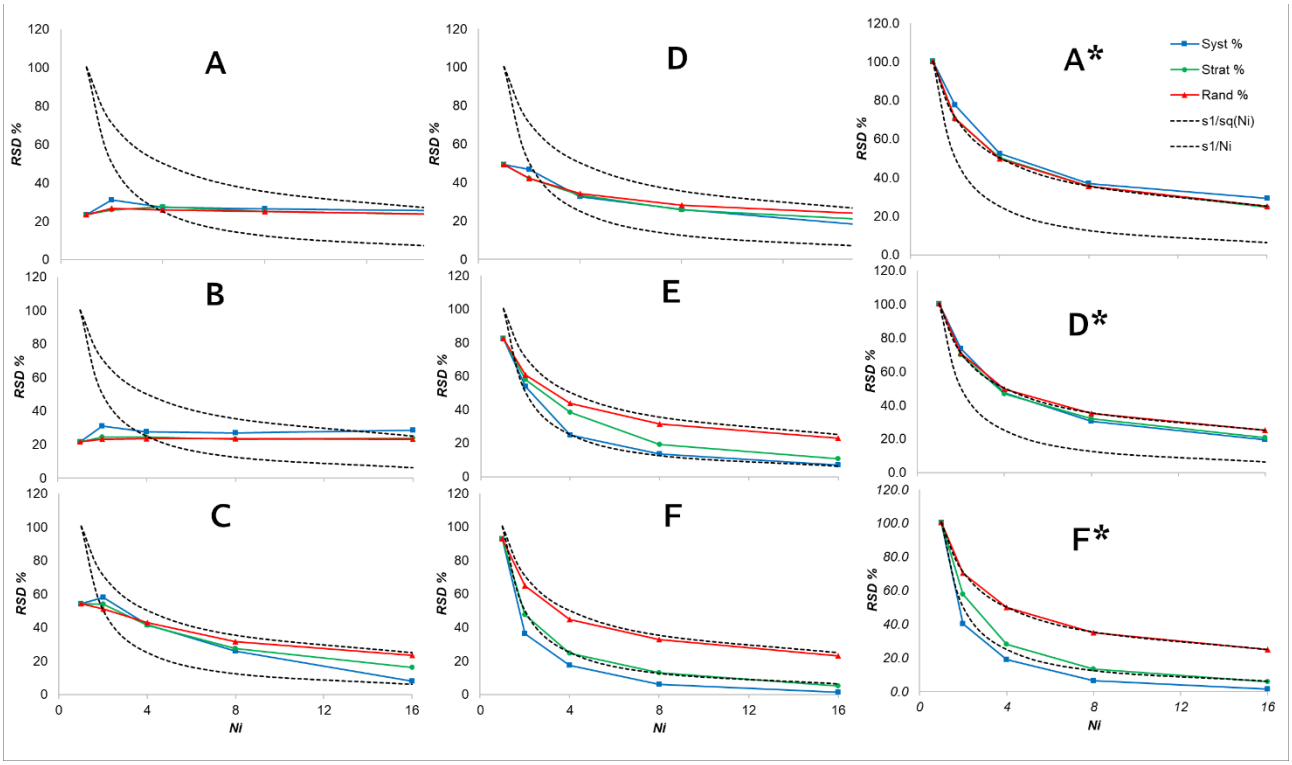

Figure S17. SU estimated for datasets A)-F) in Figure S15 calculated for a sampling *ratio* = 8. A*), D*) and F*) is for point selection, i.e. for M_{sample} << M_{Lot}. *RSD %* in % of the relative standard deviation of the lot.

In case F, where there is strong autocorrelation, SU decreases roughly proportional to 1/*Ni* for random systematic and stratified random sampling while random sampling uncertainty decreases proportional to 1/sqrt(*Ni*). For intermediate levels of autocorrelation, case C to E, the results are intermediate between case A and F.

For case A with random variations, the concentration in the mathematical samples will be the average of *jmax*·*ni* individual boxes no matter if it is single increment sampling or composite sampling. Since the distribution is random the standard deviation of the mean will be the standard deviation of the individual boxes divided by the square root of the number of boxes (*jmax*·*ni*).

It may seem strange that the uncertainty for case A and B depends on the number of data *jmax* in each increment. Suppose that the table for the lot were 512 instead of 128 points – then *jmax*·*ni* would increase 4 times and the standard deviation would decrease 2 times, when the same standard deviation was used to generate the data. The explanation is that using the same standard deviation for 128 and 512 points does *not* correspond to the same standard deviation in the lot. To represent the same concentrations in the lot, 4 points in the 512 points long dataset should be averaged to give 128 points, thus the standard deviation for 128 points would be 0.5 times the standard deviation for 512 points. And 0.5 x 2 is 1, so both lengths give the same results.

An overview of the effect of sampling ratio and number of increments on random systematic sampling is given in Fig. S18. It is obvious that the effect of the number of increments increases with the sampling ratio (from bottom to top) and with increasing autocorrelation (from A to F).

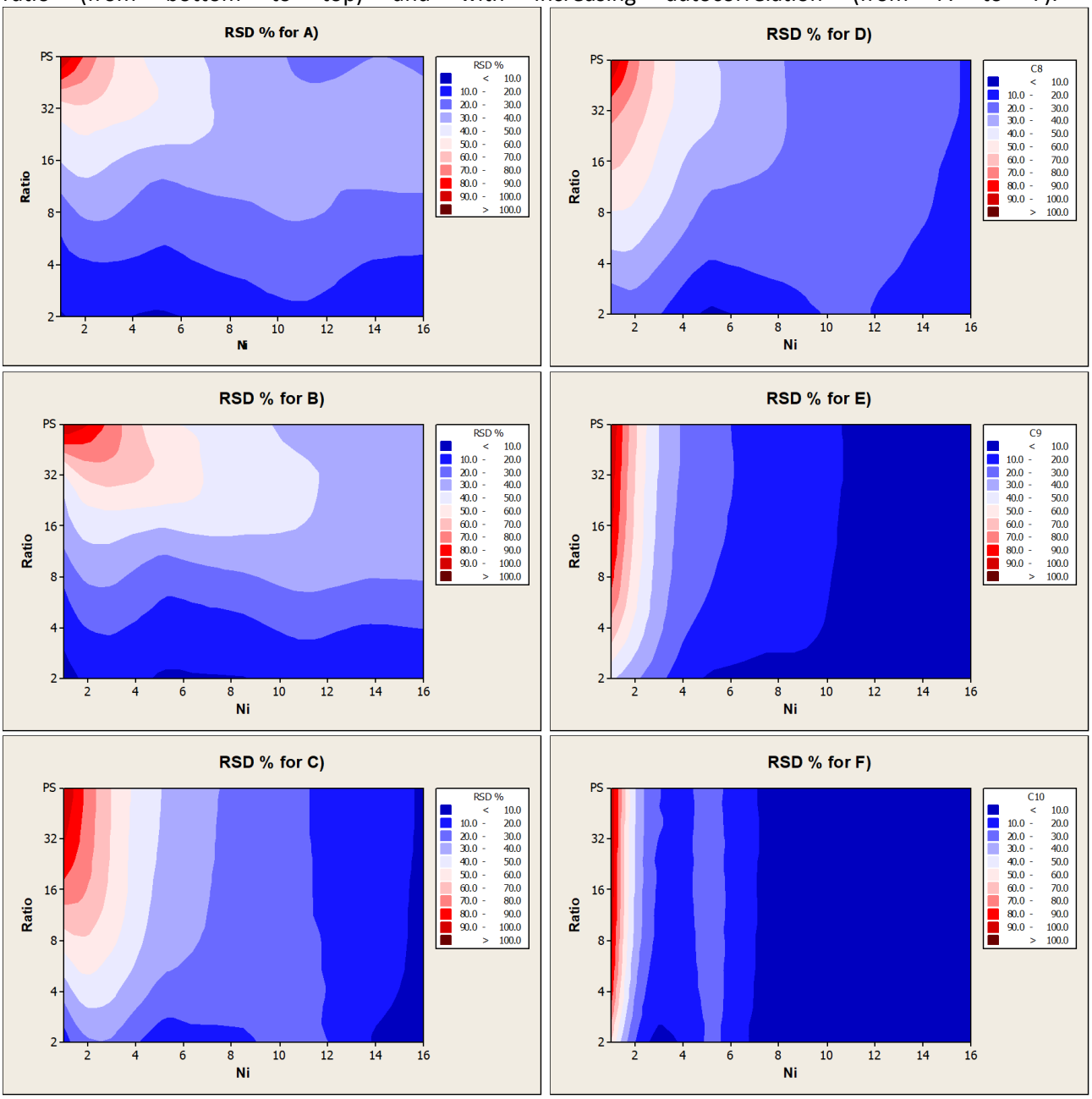

Figure S18. Predicted relative standard deviation for random systematic sampling *RSD %* in % of the standard deviation of the lot. Case A) to F) see Fig.S15. Red colors: *RSD %* < 50%. Blue colors: *RSD %*: > 50 %. PS means point selection.

The conclusion for sampling with a finite sampling ratio, i.e. for sampling where a substantial part of the lot is sampled, is that for low autocorrelation there is no real best method and the number of increments is of small importance, so decreasing the heterogeneity of the lot is the recommended tool here. For smooth distributions with high autocorrelation, random systematic or stratified random sampling with many increment is clearly the best strategy.

For point selection, i.e. for primary sampling, where the mass of the sample is negligible relative to the mass of the lot, graph A*) - F*) in Fig.17, the sampling uncertainty with data with low autocorrelation A*) will decrease roughly proportional to the square root of *Ni* and will be similar for all three methods, i.e. only the number of increments is important. For data with high autocorrelation, random systematic sampling will tend to decrease in a manner directly proportional to *Ni* or better, while stratified random sampling will have a slightly bigger uncertainty, and random sampling will still only decay proportional to the square root of *Ni*. So here random systematic or stratified random sampling with many increments is by far the best option.

S9 Small cyclic variations and the Discrete Fourier Transform DFT

Effect of small cyclic variations: Random systematic sampling is in most cases slightly better than stratified random sampling, but there is a risk if there are cyclic variations. In case of cyclic or periodic variations, random systematic sampling must be with an interval between increments of less than or equal to half the length of the cyclic period otherwise there will be aliasing of the cyclic variations (the Nyquist sampling theorem). Unfortunately, many more or less random concentration profiles contains a periodic component even without a distinct sine-wave component. For this reason stratified random sampling will in most cases be the best method. It is evident that even if the cycles cannot be seen directly in the data, they have a strong influence on the uncertainty for random systematic sampling, Fig. S19.

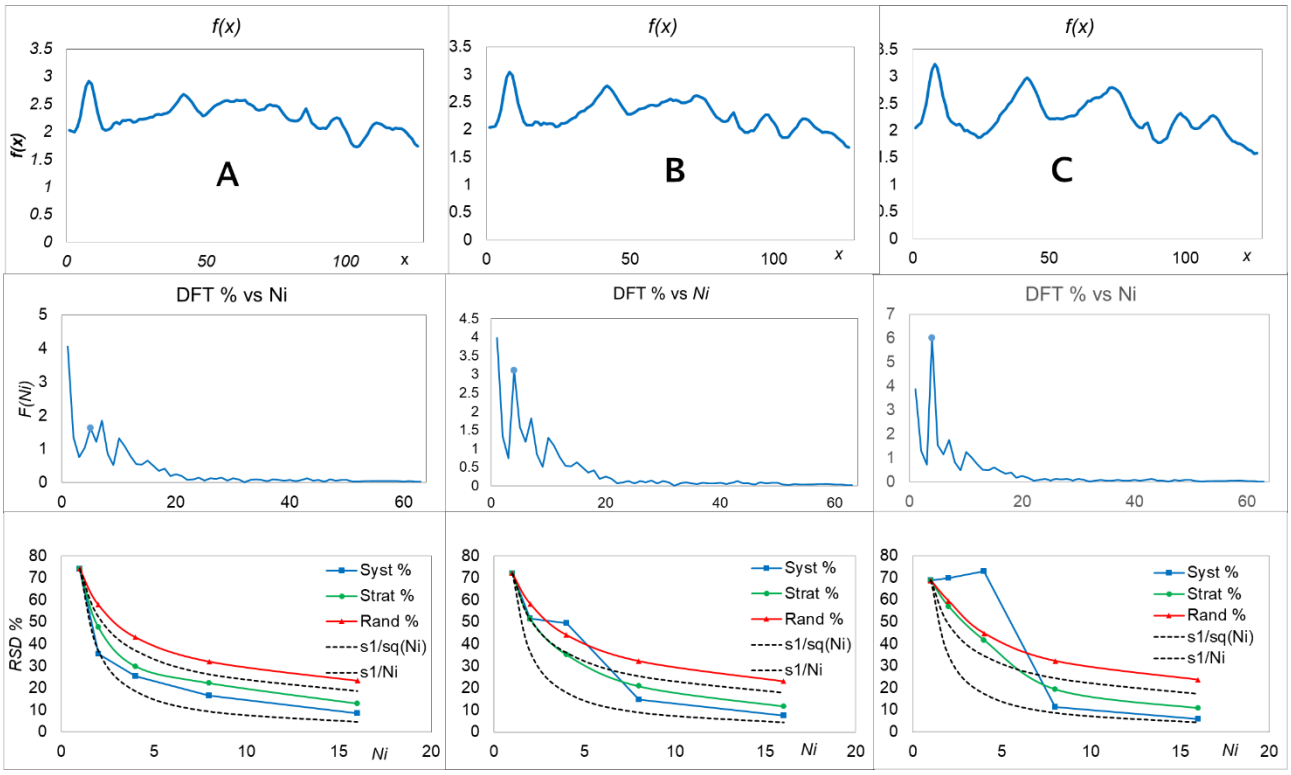

Figure S19. Effect of cyclic variations with a frequency of 4 cycles/data length, sampling *ratio* = 8. Upper: concentrations; Middle: DFT spectrum of *f(x)*, *Ni* = 4 is marked; Lower: predicted *RSD %* A) contribution for 4 cycles divided by 3; B) Unfiltered data; C) contribution for 4 cycles multiplied by 2.

Looking at Fig. S18 top, it is difficult to see that the data in case B and C contains cyclic variations. But the effect of the cycles is seen as a very high sampling uncertainty for random systematic sampling. The reason why no increased uncertainty is seen for the two peaks in the DFT spectrum to the right of the marked

peak is that they occur for *Ni* = 7 and 10, sampling intervals which is not a divisor in 128 and is not estimated in SU.

An interesting observation is that the curve for random systematic sampling with a sampling *ratio* = 2 has the same shape as the Discrete Fourier Transform of *f(x),* Fig. S20.

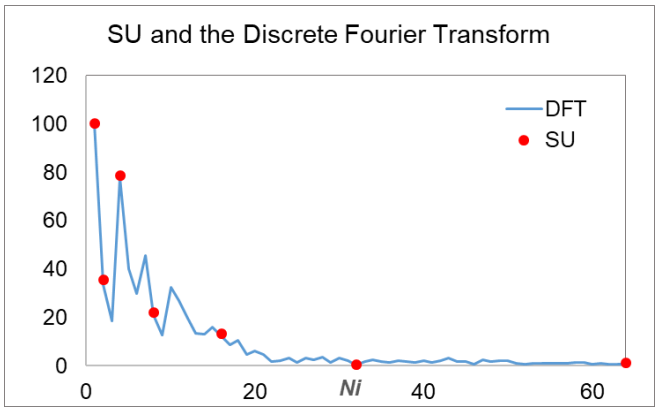

Figure S20. Comparison of the Discrete Fourier Transform of *f(x)* and SU for random systematic sampling with a *ratio* = 2, data from Fig. 16B). Data are scaled to 100 for the highest value in both series. SU is only calculated for the number of increments that are a divisor in 128.

Fig. S20 shows that there is a very good agreement between SU and DFT meaning that the high standard deviation for *Ni* = 4 in Fig. S16 is not an artefact, but an effect of cyclic variations in the data. Other cases give similar plots (for *ratio* = 2).

S10 Workbooks for prediction of SU, FSU and comparison to variograms

SU_3.0.xlsm:	Prediction of SU for point selection and sampling in general
SU-2D_3.0.xlsm:	Predictions of SU for 2-dimensional sampling
SU_Vario_3.0.xlsm:	Comparison og SU and variograms
CSU_3.0.xlsm:	Estimation of the Correct Sampling Uncertainty $CSU^2 = FSU^2 + SU^2$

New versions with numbers higher than 3.0 may be available later on.

The workbooks contain examples with data and results for Figure 7, 12, 18 and Table 1.

Manuals for use of the workbooks are given in each sheet to the right of the main tables at the top of each sheet.

The workbooks are somewhat complicated, so the best advice is to repeat the examples and change one or a few parameters at a time before using the spreadsheets for new data.