Statistical approaches for spatial sample survey: Persistent misconceptions and new developments

Article *in* European Journal of Soil Science · May 2020

DOI: 10.1111/ejss.12988

1 author:

Dick J Brus
Wageningen University and Research, Nanjing Normal University (China)
252 PUBLICATIONS **5,068** CITATIONS

SEE PROFILE

Received: 16 December 2019 | Revised: 1 May 2020 | Accepted: 8 May 2020

DOI: 10.1111/ejss.12988

REVIEW

European Journal of Soil Science WILEY

Statistical approaches for spatial sample survey: Persistent misconceptions and new developments

Dick J. Brus

Biometris, Wageningen University and Research, Wageningen, The Netherlands

Correspondence
Dick J. Brus, Wageningen University and Research, PO Box 16, 6700 AA Wageningen, the Netherlands.
Email: dick.brus@wur.nl

Abstract

Several misconceptions about the design-based approach for sampling and statistical inference, based on classical sampling theory, seem to be quite persistent. These misconceptions are the result of confusion about basic statistical concepts such as independence, expectation, and bias and variance of estimators or predictors. These concepts have a different meaning in the design-based and model-based approach, because they consider different sources of randomness. Also, a population mean is still often confused with a model mean, and a population variance with a model-variance, leading to invalid formulas for the variance of an estimator of the population mean. In this paper the fundamental differences between these two approaches are illustrated with simulations, so that hopefully more pedometricians get a better understanding of this subject. An overview is presented of how in the design-based approach we can make use of knowledge of the spatial structure of the study variable. In the second part, new developments in both the design-based and model-based approach are described that try to combine the strengths of the two approaches.

Highlights

- Ignorance of fundamental differences between design-based and model-based approaches still cause errors in statistical inference.
- Basic statistical concepts such as independence, variance and bias of an estimator have a different meaning in the two approaches.
- In estimating and testing it is important to distinguish population parameters from model parameters.
- Hybrid methods that combine the strengths of the two approaches are reviewed.

KEYWORDS

design-based approach, design independence, effective sample size, model-assisted approach, model-based approach

1 | INTRODUCTION

In 1990, de Gruijter and ter Braak (1990) published their ground-breaking paper about the fundamental difference between the design-based approach for spatial sampling and inference, based on classical sampling theory, and the model-based approach, based on geostatistical theory. Until that time many papers had been published in which the design-based approach was discarded, because if people use a probabilistic sampling scheme but observe spatial autocorrelation, they believe the independence assumption is violated. de Gruijter and ter Braak (1990) made perfectly clear that this is a misconception, that in the design-based approach no such assumption is made, and that both sampling approaches are valid and have their merits. Since then a couple of papers have been published in the soil science literature re-emphasising this important change of view, such as the paper of Papritz and Webster (1995) focusing on soil monitoring, and the discussion paper of Brus and de Gruijter (1997). According to de Gruijter, Brus, Bierkens, and Knotters (2006) the choice between the design-based and model-based approach is the most important decision to be taken in designing sampling schemes. The chapters "Global Quantities in Space" and "Local Quantities in Space" in this book (i.e. de Gruijter, Brus, Bierkens, and Knotters, 2006) have separate sections on design-based methods and model-based methods.

However, regrettably, since then also quite a few publications appeared in which the old misconceptions popped up again. Among these are papers of widely acknowledged spatial statisticians, so that there is a serious risk that the more applied pedometricians get confused again. This risk was the main motivation of this paper. A second reason for writing this paper is that there are several new developments in spatial sample survey, in both the design-based and model-based approach, trying to combine the strengths of the two approaches, which seem to be unnoticed by many pedometricians until now. The aim of this paper, therefore, is to unravel once more the misconceptions about the design-based approach and to draw the attention of pedometricians to new developments in this area.

2 | PERSISTENT MISCONCEPTIONS

2.1 | Design-based versus model-based approach

In my classes about spatial sampling I ask the participants the following question: Suppose we have measurements of a soil property, for instance soil organic matter (SOM) content, at two locations with a separation distance of 20 cm. Do you think these two measurements are correlated? I ask them to vote for one of three answers:

1 Yes, they are (> 80% confident).
2 No, they are not (>80% confident).
3 I do not know.

Most students vote for answer 1, the other students vote for answer 3, and nearly no one votes for answer 2. Then I explain that none of the answers are correct, simply because for correlation we need two series of data, not just two numbers. The question then is how to generate these two series of data. We need some random process for this. This random process differs between the design-based and model-based approach.

In the design-based approach the random process is the random selection of sampling units (in many cases these units are points), whereas in the model-based approach randomness is introduced via the statistical model of the spatial variation (Table 1). So, the design-based approach requires probability sampling, that is random sampling, using a random number generator, in such a way that all population units have a known, positive probability of being selected (Särndal, Swensson, & Wretman, 1992). A probability sampling design can be used to generate an infinite number of samples in theory, although in practical applications only one is selected for sampling.

The spatial variation model used in the model-based approach contains two terms, one for the mean (deterministic part), and an error, with a specified probability distribution. For instance, the model used in simple and ordinary kriging is (Webster & Oliver, 2007):

$$\begin{aligned} Z(\mathbf{s}) &= \mu + \epsilon(\mathbf{s}) \\ \epsilon(\mathbf{s}) &\sim \mathcal{N}\left(0, \sigma^2\right) \\ \mathrm{Cov}\{\epsilon(\mathbf{s}), \epsilon(\mathbf{s}')\} &= C(\mathbf{h}) \ , \end{aligned} \tag{1}$$

with $Z(\mathbf{s})$ the study variable at location $\mathbf{s}$, μ the constant mean (assumed to be known in simple kriging, but unknown in ordinary kriging), independent of the location $\mathbf{s}$, $\epsilon(\mathbf{s})$ the error or residual (difference between study

TABLE 1 Design-based approach versus model-based approach for sampling and statistical inference

	Sampling	Statistical inference
Design-based	Probability sampling	Design-based (no model used)
Model-based	No requirement	Based on a statistical model

variable z and mean μ) at location $\mathbf{s}$, and $C(\mathbf{h})$ the covariance of ϵ at two points separated by vector $\mathbf{h} = \mathbf{s} - \mathbf{s}'$. This model can be used to simulate an infinite number of spatial populations. All these populations together are referred to as a superpopulation (Lohr, 1999; Särndal et al., 1992). Depending on the model of spatial variation, the simulated populations may show spatial structure, because the mean is a function of covariates, as in kriging with an external drift, and/or when the errors are spatially autocorrelated. A superpopulation is a construct as the populations do not exist in the real world. The populations are similar, but not identical. For instance, the mean differs among the populations. The expectation of the population mean, that is the average over all possible simulated populations, equals the superpopulation mean, commonly referred to as the model mean, parameter μ in Eqn. 1. The variance also differs among the populations. Contrary to the mean, the average of the population variance over all populations generally is not equal to the model-variance, parameter σ^2 in Eqn. 1, but smaller. I will come to this later. The differences between the simulated spatial populations, also referred to as fields (see bottom row of Figure 1), illustrate our uncertainty about the spatial variation of the study variable in the population that is sampled or will be sampled.

In the design-based approach only one population is considered, the one sampled, but all samples that can be generated by a probability sampling design are considered. The top row of Figure 1 shows five simple random samples of size 10. The population is the same in all plots. In contrast, in the model-based approach only one sample is considered, but all populations that can be generated with the spatial variation model. The bottom row of Figure 1 shows a spatial coverage sample, superimposed on five different populations simulated with an ordinary kriging model, using a model mean of 5 mg/kg, and a spherical variogram with a nugget of 0.1, partial sill of 0.6 and a range of 75 m. Note that in the model-based approach there is no need to select a probability sample (see Table 1); there are no requirements on how the units are selected. Design-based adepts do not like to consider other populations than the one sampled, whereas model-based adepts do not like to consider other samples than the one selected. Their challenge is to get the most out of the sample that is selected.

As stressed by Brus and de Gruijter (1997), both approaches have their strengths and weaknesses. Broadly speaking, the design-based approach is most appropriate if interest is in the population mean or the population means of a restricted number of subpopulations (subareas). The model-based approach is most appropriate if our aim is to map the soil property of interest. Further, the strength of the design-based approach is the validity of the estimates. Validity means that an objective assessment of the uncertainty of the estimator is warranted,

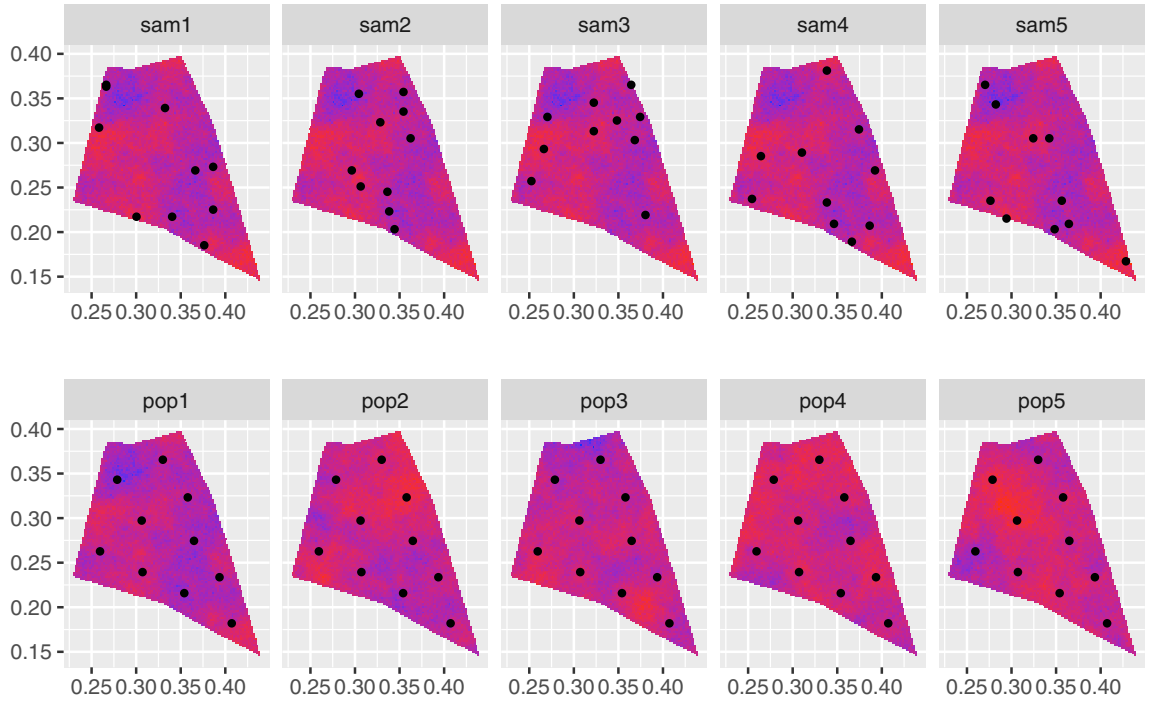

FIGURE 1 Random process considered in the design-based approach (top row) and model-based approach (bottom row). In the design-based approach only the sampled population is considered, but all samples that can be generated by the sampling design. In the model-based approach all populations that can be generated by the model are considered, but only the sample that is selected. The populations represent organic matter concentration in g kg^{-1} in the topsoil. The coordinates along the horizontal and vertical axis are in km

and that the coverage of confidence intervals is almost correct, provided that the sample is large enough to assume an approximately normal distribution of the estimator and design-unbiasedness of the variance estimator (Särndal et al., 1992). The strength of the model-based approach is efficiency, that is, more precise estimates of the (sub)population mean given the sample size, provided that a valid model is used. However, this requires stringent checks of the validity of the model. If one wishes to be on the safe side about the validity of the results, a design-based approach is the best choice, as no modelling assumptions are made.

2.2 | i.i.d

In a recent review paper on spatial sampling by Wang, Stein, Gao, and Ge (2012) there is a section with the caption "Sampling of i.i.d populations." i.i.d. stands for identically and independently distributed. In this section by Wang et al. (2012) we can read: "In SRS (simple random sampling, DB) it is assumed that the population is independent and identically distributed". This is one of the old misconceptions revitalized by this review paper. I will make clear that in statistics i.i.d is not a characteristic of populations, so the concept of i.i.d populations does not make sense. The same misconception can be found in Plant (2012, p.147): "There is considerable literature on sample size estimation, much of which is discussed by Cochran (1977, chapter 4). This literature, however, is valid for samples of independent data but may not retain its validity for spatial data". Also, according to Wang, Haining, and Cao (2010) the classical formula for the variance of the estimated mean with simple random sampling, $V = \sigma^2/n$, only holds when data are independent. They say: "However in the case of spatial data, although members of the sample are independent by construction, data values that are near to one another in space, are unlikely to be independent because of a fundamental property of attributes in space, which is that they show spatial structure or continuity (spatial autocorrelation)". According to Wang et al. (2010) the variance should be approximated by:

$$V(\hat{\bar{z}}) = \sigma^2 - \frac{\overline{\mathrm{Cov}(z_i, z_j)}}{n}, \tag{2}$$

with $V(\hat{\bar{z}})$ the variance of the estimated regional mean (mean of spatial population), σ^2 the population variance, n the sample size, and $\overline{\mathrm{Cov}(z_i, z_j)}$ the average autocovariance between all pairs of individuals (i, j) in the population (sampled and unsampled). So according to this formula, ignoring the mean covariance within the population leads to an overestimation of the variance of the estimated mean. In Section 2.4 I will make clear that this formula is incorrect, and that the classical formula is still valid, also for populations showing spatial structure or continuity.

Remarkably, in other publications we can read that the classical formula for the variance of the estimated population mean with SRS *underestimates* the true variance for populations showing spatial structure (see for instance Griffith (2005) and Plant (2012)). The reasoning is that due to the spatial structure there is less information in the sample data about the population mean. In Section 2.4 I explain that this is also a misconception. Do not get confused by these publications, and stick to the classical formulas, which you can find in standard textbooks on sampling theory such as Cochran (1977) and Lohr (1999). For instance, for simple random sampling with replacement from finite populations the variance of the estimated mean equals:

$$V(\hat{\bar{z}}) = \frac{S^2}{n}, \tag{3}$$

with S^2 the population variance:

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})^2, \tag{4}$$

with N the total number of population units and $\bar{z}$ the population mean. An estimator of the variance of the estimated mean is obtained by replacing the population variance S^2 in Eqn. 3 by its estimator:

$$\hat{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \hat{\bar{z}})^2, \tag{5}$$

with $\hat{\bar{z}}$ the estimated population mean, estimated by the sample average. Eqn. 3 also holds for simple random sampling from infinite populations.

The concept of independence of random variables is illustrated with a simulation. The top row of Figure 2 shows five SRSs of size two. The two points are repeatedly selected from the same population (showing clear spatial structure), so this top row represents the design-based approach. The bottom row shows two points, not selected randomly and independently, but at a fixed distance of 10 m. These two points are placed on different populations generated by the model described above, so the bottom row represents the model-based approach. The values measured at the two points are plotted against each other in a scatter plot, but now not for just five SRSs or five populations, but for 1,000 samples and populations (Figure 3). As we can

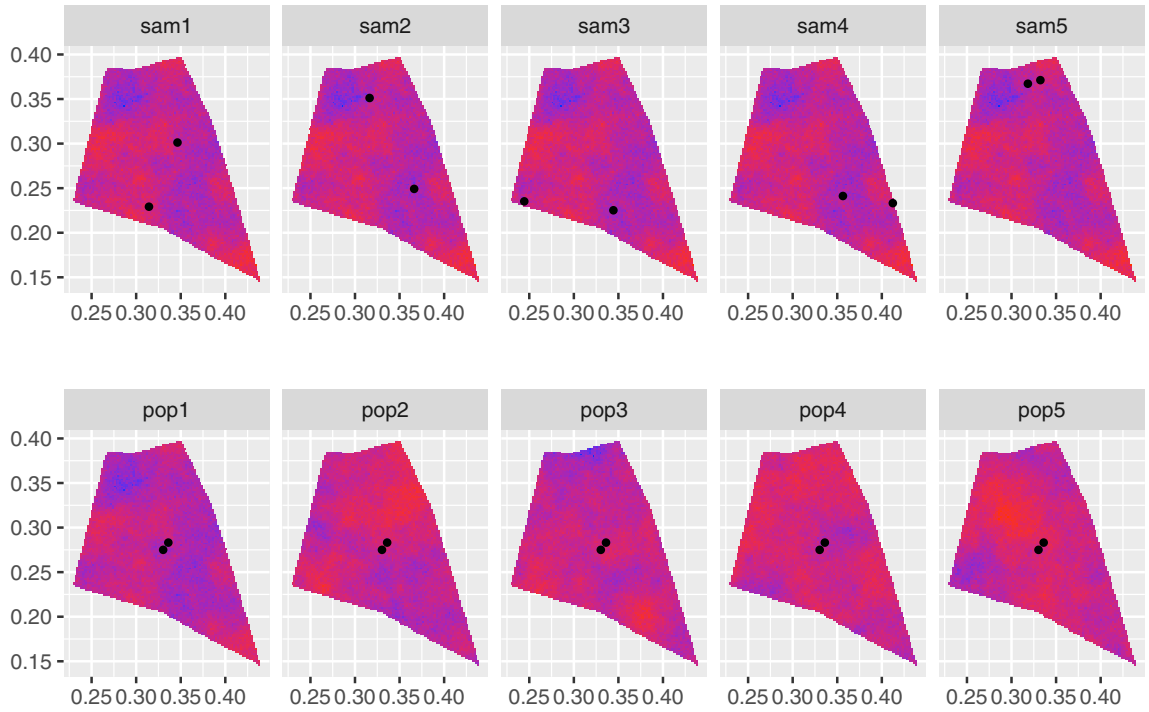

FIGURE 2 Illustration of independence in design-based and model-based approaches. The top row shows five samples of two points selected randomly and independently from each other from one population (design-based approach). The bottom row shows two points not selected randomly, at a distance of 10 m from each other from five model realizations (model-based approach)

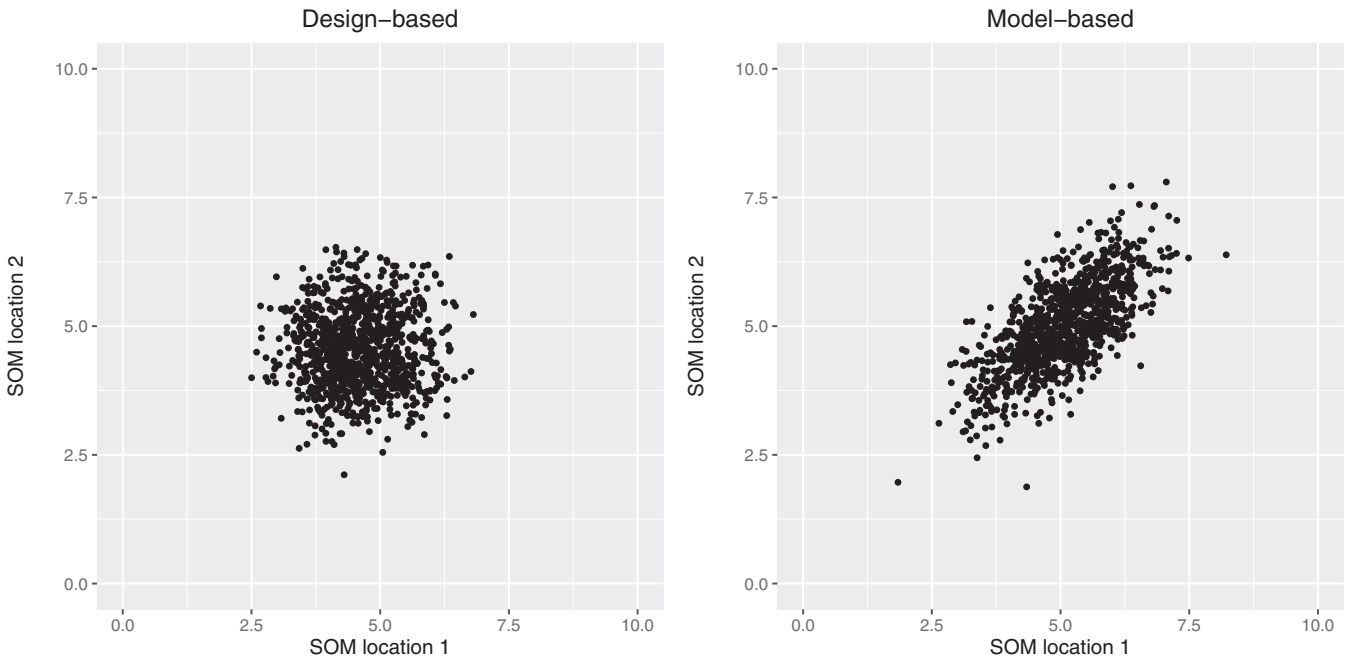

FIGURE 3 Scatter plot of the values at two randomly and independently selected points, 1,000 times selected from one population (design-based approach), and at two fixed points with a separation distance of 10 m, selected non-randomly from 1,000 model realizations (model-based approach). Figure 2 shows the first five pairs of points for each sampling approach. SOM: soil organic matter concentration (g/kg)

see, there is no correlation between the two variables generated by the repeated random selection of the two points, whereas the two variables generated by the repeated simulation of populations are correlated.

Instead of two points, we may select two series of probability samples independently from each other, for instance, two series of simple random samples of size 10, or two series of systematic random samples with

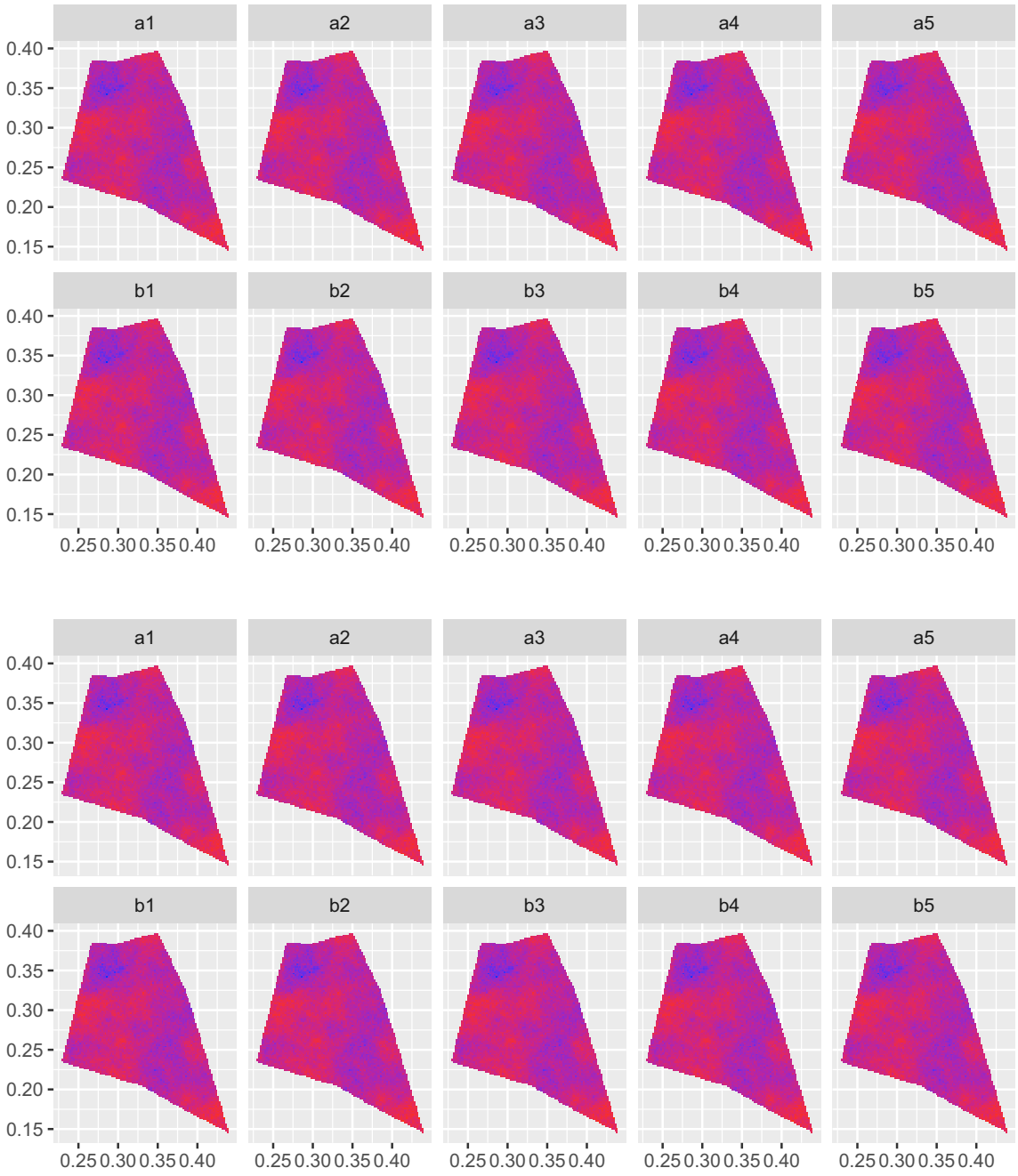

FIGURE 4 Two series (a and b) of simple random samples of 10 points (top), and two series of systematic random samples of 10 points on average (bottom). The samples of series a and b are selected independently from each other

random origin with an average size of 10 (see Figure 4). Again, if we plot the sample averages of pairs of simple random samples (SRS) and pairs of systematic random samples (SY) against each other, we see that the two averages are not correlated (Figure 5). Note that the variation of the averages of the SY samples is considerably smaller than that of the SRS samples. The sampled population shows spatial structure, and by spreading the locations out over the spatial population, the precision of the estimated population mean is increased.

This sampling experiment shows that independence is not a characteristic of a population, as stated by Wang et al. (2012), but of random variables (in the experiment the values at points, or the sample averages) generated by a random process. As the random process differs between the design-based and model-based approaches, independence has a different meaning in these two approaches. For that reason, it is recommended to be more specific when using the term independence, by saying that data *are design-independent* or that you *assume* that the data are *model-independent*.

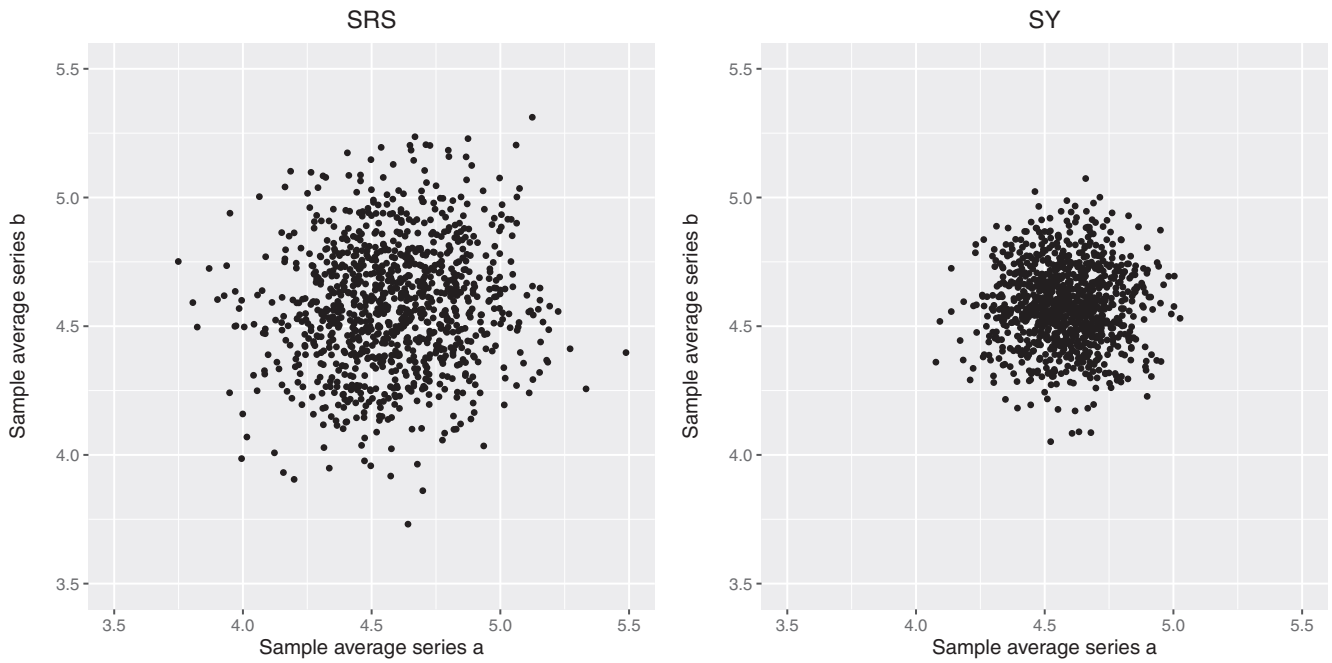

FIGURE 5 Scatterplot of averages of 1,000 pairs of simple random samples of 10 points, and of averages of 1,000 pairs of systematic random samples of 10 points on average. Figure 4 shows the first five pairs for simple random sampling (SRS) and systematic random sampling (SY)

2.2.1 | Superpopulation versus sampling model

Not all models presented in publications are superpopulation models. Some of them are sampling models. A sampling model describes the random process generating the sample data, not the random process that may have generated the entire population. For instance, the model

$$\begin{aligned} z(\mathbf{s}) &= \mu + \epsilon(\mathbf{s}) \\ \epsilon(\mathbf{s}) &\overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma^2\right) , \end{aligned} \tag{6}$$

at first sight may look to be a superpopulation model, assuming a constant model mean and model-variance, and model-independent data (no spatial autocorrelation). However, it may appear that μ and σ are not defined as model parameters, but population parameters, and that the model describes the random process generating sample data through simple random sampling. This implies that i.i.d stands here for design-independent. To avoid confusion, I would rewrite this model as follows:

$$\begin{aligned} z(\mathbf{S}) &= \bar{z} + \epsilon(\mathbf{S}) \\ \epsilon(\mathbf{S}) &\overset{i.i.d.}{\sim} \mathcal{N}\left(0, S^2\right) . \end{aligned} \tag{7}$$

Note that I use capital **S** to indicate that the locations are random variables, and as a consequence the residual ϵ and the study variable z at a random location **S** are also random variables. Further note that the normal distribution here refers to the frequency distribution of the data (residuals) in the sampled population of interest.

I think the book by Lohr (1999) is somewhat confusing on this aspect. At the end of each chapter in which a probability sampling design type is described, there is a section on a model for that sampling design. For instance, in Section 4.6 "A model for Stratified Sampling", she says that a model for this sampling design is an ANOVA model with fixed effects. In this model it is assumed that the population consists of several groups (subpopulations), that all units within the same group have the same expected value and variance, and that the model covariance is zero. In formula, for $g = 1, \cdots, G$,

$$\begin{aligned} E\{y(\mathbf{s})\} &= \beta_g \\ V\{y(\mathbf{s})\} &= \sigma_g^2 \\ C\{y(\mathbf{s}), y(\mathbf{s}')\} &= 0 , \end{aligned} \tag{8}$$

for all **s** and $\mathbf{s}' \in U_g$, with β_g and σ_g^2 the model mean and model-variance of group g, and U_g group g. The covariance of the random variables in different groups is also 0. Although at the end of this section she stresses: "If a different model is used, however, then different estimates are obtained." However, if a stratified random sample is selected, in design-based estimation of the mean we do not *assume* that the population is generated by an ANOVA model. No modelling assumptions are made in

the design-based approach. In a model-based approach, the spatial structure in the population is possibly better modelled by a linear mixed model with a spatial random effect, in which the mean is a function of a categorical predictor (used as strata) and the residuals are spatially correlated. However, by selecting the points independently from each other, the data are design-independent.

2.3 | Bias and variance

Bias and variance are commonly used statistics to quantify the quality of an estimator. Bias quantifies the systematic error, variance the random error of the estimator. Both are defined as expectations. But expectations over the realizations of which random process? Over realizations of a probability sampling design (samples) or realizations of statistical model (populations)? Like independence, it is important to distinguish *design-bias* from *model-bias*, and *design-variance* (commonly referred to as sampling variance) from *model-variance*.

The concept of model-unbiasedness deserves more attention. Figure 6 shows a preferential sample from a population simulated by sequential Gaussian simulation with a constant mean of 10 and an exponential variogram without nugget, a sill of 5 and a distance parameter of 20. The points are selected by sampling with draw-by-draw selection probabilities proportional to size (pps-sampling) (de Gruijter et al., 2006), using the square of the simulated values as a size variable. We may have a similar sample that is collected for delineating soil contamination or detecting hot spots of soil bacteria, etc. Many samples are selected at locations with a large value, few points at locations with a small value. The sample data are used in ordinary kriging (Figure 6). The prediction errors are computed by subtracting the kriged map from the simulated population. Figure 7 shows a histogram of the prediction errors. The population mean error equals 0.483, not zero. You may have expected a positive systematic error because of the over-representation of locations with large values, but on the other hand, kriging predictions are best linear unbiased predictions (BLUP), so from that point of view, this systematic error might be unexpected. BLUP means that at individual locations the ordinary kriging predictions are unbiased. However, apparently this does not guarantee that the average of the prediction errors, averaged over all population units, equals zero. The reason is that unbiasedness is defined here over all realizations (populations) of the statistical model of spatial variation. So, the U in BLUP stands for model-unbiasedness. For other model realizations, sampled at the same points, we may have much smaller values, leading to a negative mean error of that population. On average, over all populations, the error at any point will be zero, and consequently also the average over all populations of the mean error.

This experiment shows that model-unbiasedness does not protect us against selection bias, that is, bias due to preferential sampling. With preferential sampling it is evident that we should respect the sampling design, also in model-based prediction (for how this can be done, see Section 3.2).

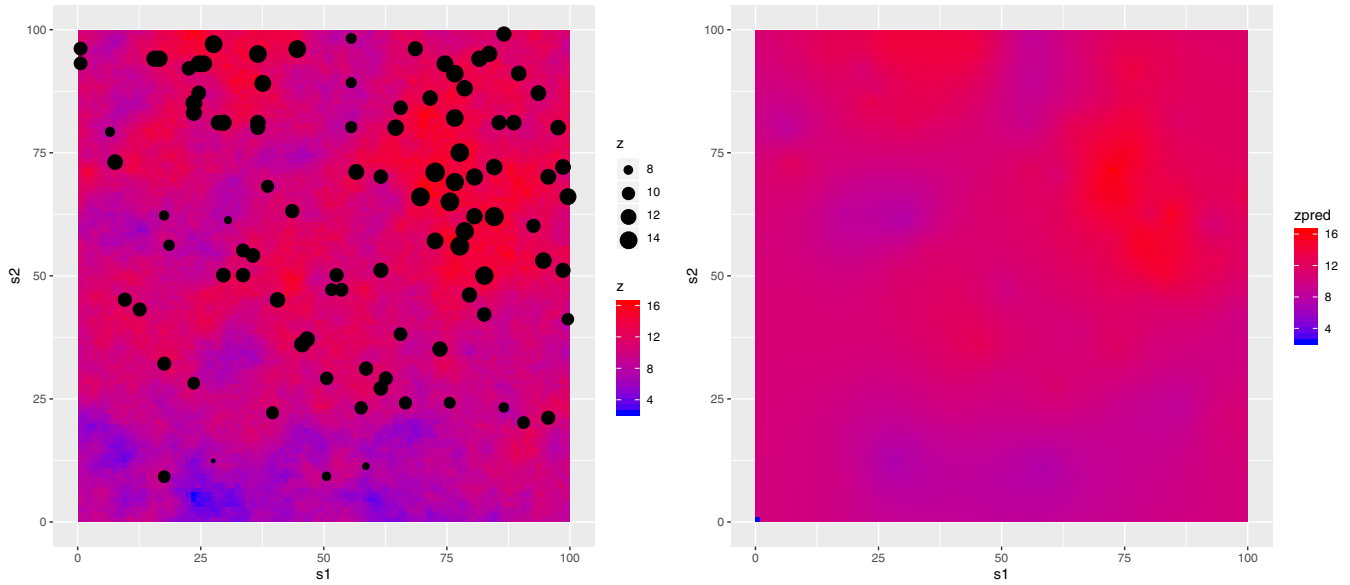

FIGURE 6 Preferential sample and ordinary kriging predictions

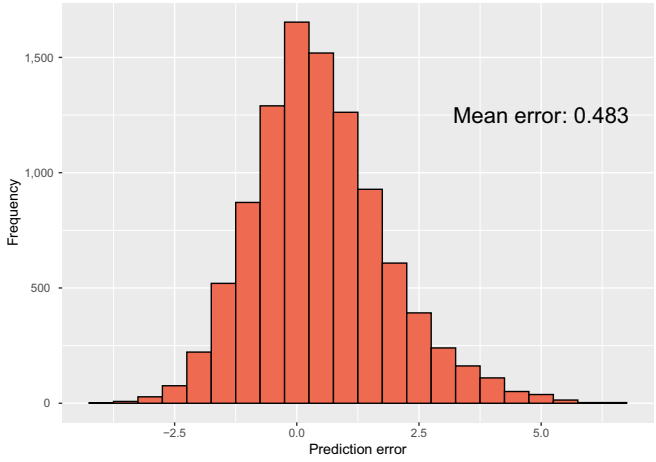

FIGURE 7 Histogram of errors of ordinary kriging predictions from a preferential sample (Figure 6)

2.4 | Effective sample size

Another persistent misconception is that when estimating the variance of the estimated mean of a spatial population or the correlation of two variables of a population we must account for autocorrelation of the sample data. This misconception occurs, for instance, in Griffith (2005) and in various sections (for instance, sections 3.5, 10.1 and 11.2) of Plant (2012). The reasoning is that, due to the spatial autocorrelation in the sample data, there is less information in the data about the parameter of interest, and so the effective sample size is smaller than the actual sample size. An early example of this misconception is Barnes' publication on the required sample size for estimating nonparametric tolerance intervals (Barnes, 1988). de Gruijter and ter Braak (1990) showed that a basic probability sampling design like simple random sampling requires fewer sampling points than the model-based sampling design proposed by Barnes.

The misconception is caused by confusing population parameters with model parameters. Recall that the population mean and the model mean are not the same; the model mean μ of Eqn. 1 is the expectation of the population means over all populations that can be simulated with the model. The same holds for the variance of a variable, and the covariance and Pearson correlation coefficient of two variables. All these parameters can be defined as a parameter of a (finite or infinite) population, or of random variables generated by a superpopulation model. Using an effective sample size to quantify the variance of an estimator is perfectly correct for model parameters, but not so for population parameters. For instance, when the correlation coefficient is defined as a population parameter and sampling locations are selected by simple random sampling, there is no need to apply the method proposed by Clifford, Richardson, and Hemon (1989) to correct the p-value in a significance test for the presence of spatial autocorrelation.

I elaborate on this for the mean as the parameter of interest. Suppose a sample is selected in some way (need not be random), and the sample average is used as an estimate of the model mean. Note that for a model with a constant mean as in Eqn. 1, the sample average is a model-unbiased estimator of the model mean, but in general not the best linear unbiased estimate (BLUE) of the model mean. If the random variables are model-independent, the variance of the sample average as an estimate of the model mean can be computed by:

$$V(\hat{\mu}) = \frac{\sigma^2}{n}, \tag{9}$$

with σ^2 the model-variance of the random variable (see Eqn. 1). The variance presented in Eqn. 9 necessarily is a model-variance as it quantifies our uncertainty about the model mean, which only exists in the model-based approach. If the random variables are not model-independent, the model-variance of the sample average can be computed by (de Gruijter et al., 2006):

$$V(\hat{\mu}) = \frac{\sigma^2}{n}\{1 + (n-1)\bar{\rho}\}, \tag{10}$$

with $\bar{\rho}$ the mean correlation within the sample (the average of the correlation of all pairs of sampling points). The term inside the curly brackets is larger than 1, unless $\bar{\rho}$ equals zero. So, the variance of the estimated model mean with dependent data is larger than when data are independent. The number of independent observations that is equivalent to a spatially autocorrelated dataset's sample size n, referred to as the effective sample size, can be computed with (de Gruijter et al., 2006):

$$n_{\text{eff}} = \frac{n}{\{1 + (n-1)\bar{\rho}\}}. \tag{11}$$

So, if we substitute n_{eff} for n in Eqn. 9, we obtain the variance presented in Eqn. 10. Equation 11 is equivalent to Eqn. 2 in Griffith (2005). Figure 8 shows that the effective sample size decreases sharply with the mean correlation. With a mean correlation of 0 the effective sample size equals the actual sample size; with a mean correlation of 1 the effective sample size equals 1.

To illustrate the difference between the model-variance and design-variance of a sample average, I simulated a finite population of 100 units, located at the nodes of a square grid, with a model mean of 10, an exponential

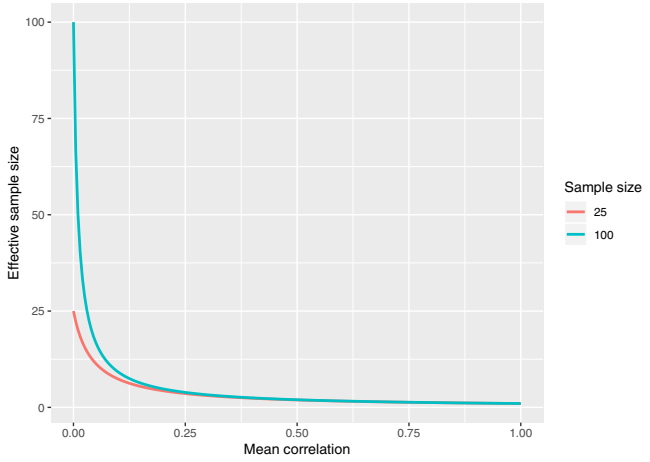

FIGURE 8 Effective sample sizes for samples of size 25 and 100, as a function of the mean correlation within the sample

variogram without nugget, an effective range of three times the distance between adjacent population units, and a sill of 1 (Figure 9). The model-variance of the average of a simple random sample *without replacement* of size n is computed using Eqn. 10. The design-variance of the sample average, used as an estimate of the population mean, is computed by:

$$V(\hat{\bar{z}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, \tag{12}$$

with N the total number of population units ($N = 100$). This is done for a range of sample sizes: $n = 10, 11, \cdots, 100$. Note that for $n < 100$ the model-variance of the sample average for a given n, differs between samples. For samples showing strong spatial clustering, the mean correlation is relatively large, and consequently the model-variance is relatively large (see Eqn. 10). There is less information in these samples about the model mean than in samples without spatial clustering of the points. Therefore, to estimate the expectation of the model-variance over repeated simple random sampling for a given n, I selected 200 simple random samples of that size n, and I averaged the 200 model-variances. Figure 10 shows the result. Both the model-variance and the design-variance of the sample average decrease with the sample size. For all sample sizes the model-variance is larger than the design-variance. The design-variance goes to zero, for $n = 100$ (see Eqn. 12), whereas the model-variance for $n = 100$ equals 0.0509. This can be explained as follows. Although with $n = 100$ we know the population mean without error, this population mean is only an estimate of the model mean. Recall that the model mean is the expectation of the population mean over all realizations of the model. In Figure 11 we can see that the population mean shows considerable variation. The variance of 10,000 simulated population means equals 0.0513, which is nearly equal to the value of 0.0509 for the model-variance computed with Eqn. 10. To conclude, it is not useful to compare the design-variance and the model-variance of the sample average, as they quantify the

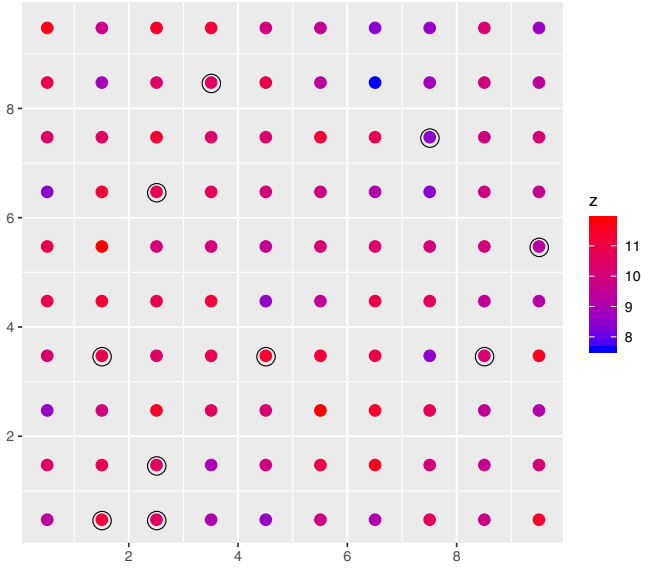

FIGURE 9 Simple random sample without replacement of 10 points from a finite population simulated with a model with a model mean of 10, model-variance of 1 and an exponential variogram (without nugget) with a distance parameter equal to the distance between neighbours (effective range is three times this distance). The mean correlation within the sample equals 0.135, and the model-variance of the estimated model mean equals 0.222

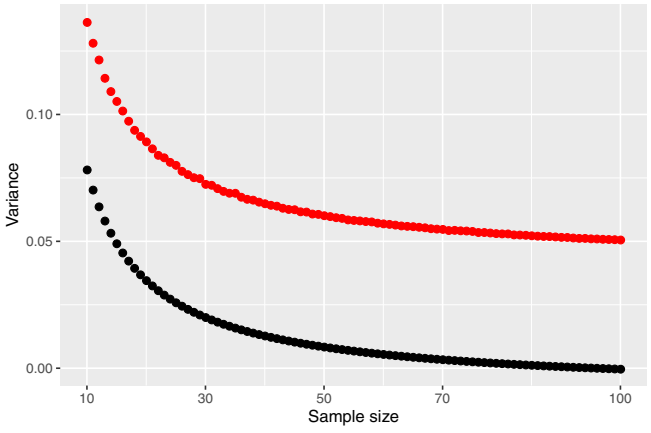

FIGURE 10 Model variance (red line) and design-variance (black line) of the average of simple random sample without replacement from finite population of Figure 9, as a function of the sample size

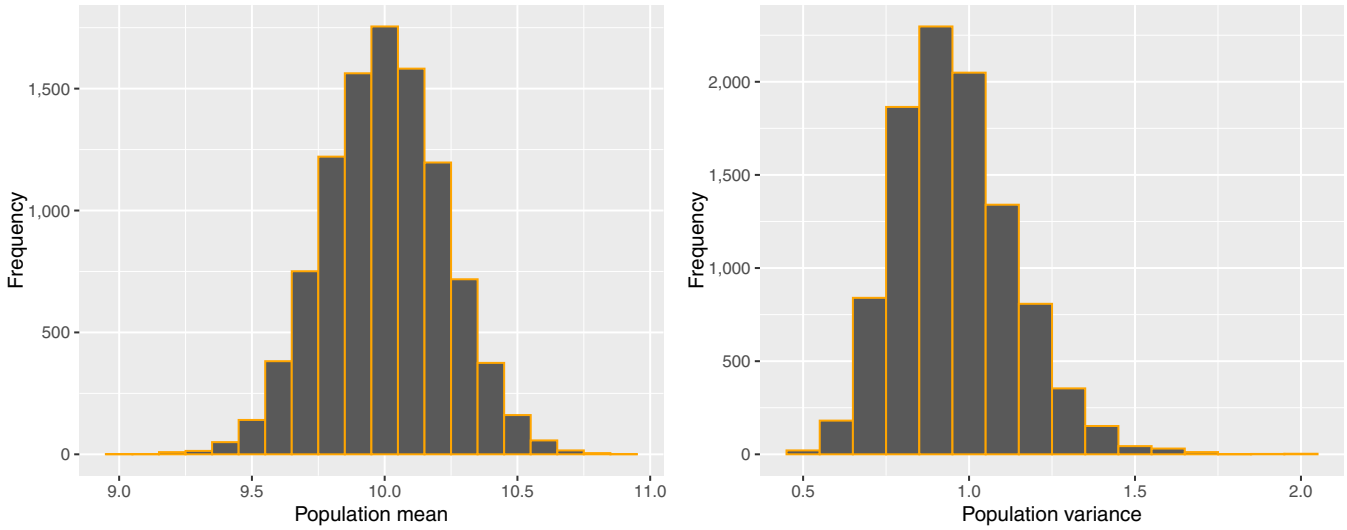

FIGURE 11 Histograms of means and variances of 10,000 simulated populations

uncertainty about different quantities, the population mean and the model mean, respectively.

In observational research I cannot think of situations in which interest is in estimation of the mean of a super-population model. This in contrast to experimental research. In experimental research we are interested in the effects of treatments; think for instance of the effects of different types of soil tillage on the soil carbon stock. These treatment effects are quantified by different model means. Also, in time-series analysis of data collected in observational studies we might be more interested in the model mean than in the mean over a bounded period of time.

Now let us return to Eqn. 2. What is wrong with this variance estimator? Where Griffith (2005) confused the population mean and the model mean, Wang et al. (2010) confused the population variance with the sill variance (*a priori* variance) of the random process that has generated the population (Webster & Oliver, 2007). The parameter σ^2 in their formula is defined as the population variance, and in doing so the variance estimator is clearly wrong. However, if we define σ^2 in this formula as the sill variance, the formula makes more sense, but even then, the equation is not fully correct. The variance computed with this equation is not the design-variance of the average of a simple random sample selected from the sampled population, but the expectation of this design-variance over all realizations of the model. So, it is a model-based prediction of the design-variance of the estimated population mean, estimated from a simple random sample. For the population actually sampled, the design-variance is either smaller or larger than this expectation. Figure 11 shows that there is considerable variation in the population variance among the 10,000 populations simulated with the model. Consequently, for an individual population the variance of the estimated population mean, estimated from a simple random sample, can largely differ from the model-expectation of this variance. Do not use Eqn. 2 for estimating the design-variance of the estimated population mean, but simply use Eqn. 12 (for SRS with replacement and SRS of infinite populations the term $(1 - n/N)$ can be dropped). Eqn. 2 is only relevant for comparing SRS under a variety of models of spatial variation (Domburg, de Gruijter, & Brus, 1994; Ripley, 1981).

2.5 | Exploiting spatial structure in a design-based approach

Another misconception is that in the design-based approach the possibilities of exploiting our knowledge about the spatial structure of the study variable are limited, because the sampling locations are selected randomly. This would indeed be a very serious drawback, but happily enough, this is not true. There are various ways of utilizing this knowledge. Our knowledge about the spatial structure can be used at the stage of designing the sample and/or at the stage of the statistical inference once the data are collected (Table 2). I distinguish two situations, one in which maps of covariates are available, one in which such maps are lacking. In the first situation, the covariate maps can be used, for instance, to stratify the population. With a quantitative covariate, optimal stratification methods are available (Baillargeon & Rivest, 2011; Brus, Yang, & Zhu, 2019; de Gruijter, Minasny, & McBratney, 2015). Other options are, for instance, pps-sampling and πps-sampling (Särndal et al., 1992), balanced sampling (Brus, 2015; Deville & Tillé, 2004) and

TABLE 2 Exploiting knowledge about the spatial structure of the study variable in the design-based approach

Stage	Covariates available	No covariates available
Sampling	Stratified random sampling	Systematic random sampling
	Pps and πps sampling	Compact geographical stratification
	Balanced sampling	Geographical spreading with LPM
	Covariate-space spreading with LPM	Generalized random tessellation stratified sampling
Inference	Model-assisted, using regression or ratio model	Model-assisted, using penalized spline regression model

Abbreviation: LPM, Local pivotal method.

well-spread sampling in covariate space with the local pivotal method (Grafström, Lundström, & Schelin, 2012). At the inference stage the covariate maps can be used in a model-assisted approach (see Section 3.1), using, for instance, a linear regression model to increase the precision of the design-based estimator.

If no covariate maps are available, we may anticipate the presence of spatial structure by spreading out the sampling locations throughout the study area. This spreading can be done in many ways, for instance by systematic random sampling, compact geographical stratification (Brus, Spätjens, & de Gruijter, 1999), well-spread sampling in geographical space with the local pivotal method (Grafström et al., 2012), and generalized random tessellation stratified sampling (Stevens & Olson, 2004). At the inference stage, again a model-assisted approach can be advantageous (see Section 3.1).

3 | NEW DEVELOPMENTS

In this section I describe several developments in spatial sample survey that combine the strengths of the two approaches. Some of these developments already have a long history in sample survey, others are more recent developments. In the design-based approach various estimators are developed that exploit a superpopulation model of the study variable, leading to more accurate estimates, while maintaining the validity of the design-based approach. The combination of probability sampling and estimators that are built on superpopulation models, is coined as the model-assisted approach.

On the other side, in the model-based approach, methods have been developed that account for differences in the selection probabilities of sampling locations, so that the systematic error in the model-based predictions is reduced. This can be seen as a move towards the design-based approach.

3.1 | Modelling spatial variation in a design-based approach

The model-assisted approach tries to build the strength of the model-based approach, a potential increase of the accuracy of estimates, into the design-based approach. As in the design-based approach, sampling units are selected by probability sampling, and consequently bias and variance are defined as design-bias and design-variance (Table 3). Like the model-based approach, a superpopulation model is used; however, as explained hereafter, the role of the model differs from its role in the model-based approach. To stress the different role in the model-assisted approach, the model is referred to as the "working model".

Breidt and Opsomer (2017) present an overview of model-assisted estimators motivated by prediction ideas. To predict the study variable a superpopulation model is used. A general formulation of this working model is:

$$z_i = \mu(\mathbf{x}_i) + \epsilon_i, \tag{13}$$

with $\mu(\mathbf{x}_i)$ the model mean (superpopulation mean) for unit i, which is a function of the values of the covariates at that location collected in vector $\mathbf{x}_i$, and ϵ_i a random variable with zero mean. In the following subsections various functions for $\mu(\mathbf{x}_i)$, both linear functions (Subsections 3.1.1 and 3.1.2), and non-linear functions (Subsection 3.1.3), are described.

If the study variable and the covariates were observed for all population units, the parameters of the function for the model could be computed from all population units. Think, for instance, of the regression coefficients of a linear regression model computed from the entire population. These parameters could then be used to compute the means $m(\mathbf{x}_i)$, that is the population-level fit of the model means $\mu(\mathbf{x}_i)$. In practice we have a sample only, which is used to estimate the means $m(\mathbf{x}_i)$ by $\hat{m}(\mathbf{x}_i)$. These estimates are plugged into the model-assisted difference estimator:

$$\hat{\bar{z}}_{\text{dif}} = \frac{1}{N}\sum_{i=1}^{N}\hat{m}(\mathbf{x}_i) + \frac{1}{N}\sum_{i=1}^{n}\frac{z_i - \hat{m}(\mathbf{x}_i)}{\pi_i}, \tag{14}$$

TABLE 3 Properties of three approaches for sampling and statistical inference

Approach	Sampling	Inference	Regression coefficients	Quality criteria
Design-based	Probability sampling	Design-based	No model	Design-bias, design-variance
Model-assisted	Probability sampling	Model-assisted	Population parameters	Design-bias, design-variance
Model-based	No requirement	Model-dependent	Superpopulation parameters	Model-bias, model-variance

with π_i the inclusion probability of sampling location i, as determined by the sampling design. The first term is the population mean of the model predictions of the study variable using the sample estimates of the model parameters, the second term is the Horvitz-Thompson estimator of the population mean of the residuals.

The variance of the model-assisted difference estimator equals the variance of the Horvitz-Thompson estimator of the population mean of the differences $d_i = z_i - m(\mathbf{x}_i)$:

$$V\left(\hat{\bar{z}}_{\text{dif}}\right) = V\left(\hat{\bar{d}}_{\text{HT}}\right) . \tag{15}$$

An estimator is obtained by substituting $\hat{m}(\mathbf{x}_i)$ for $m(\mathbf{x}_i)$ to compute the differences.

A wide variety of model-assisted estimators have been developed and tested in the past decades. They differ in the working model used to obtain the estimates $\hat{m}(\mathbf{x}_i)$. The best-known class of model-assisted estimator is the generalized regression estimator (Särndal et al., 1992). Alternative model-assisted estimators with potentials for spatial sample survey are, for instance, the model-assisted estimators based on a penalized spline regression model and on machine learning techniques.

3.1.1 | Generalized regression estimator

The working model of the generalized regression estimator is the heteroscedastic multiple regression model:

$$z_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \tag{16}$$

with ϵ_i uncorrelated, zero mean, and model-variance σ_i^2. So, $\mu(\mathbf{x}_i)$ in Eqn. 11 is $\mathbf{x}_i{}^{\mathrm{T}}\boldsymbol{\beta}$. If $\{z_i, x_{1,\,i}, \cdots, x_{J,\,i}\}$ (J is number of covariates +1) were observed for all units $i = 1, \cdots, N$ in the population, the regression coefficients $\boldsymbol{\beta}$ would be estimated by:

$$\mathbf{B} = \left(\sum_{i=1}^{N}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\sigma_i^2}\right)^{-1}\sum_{i=1}^{N}\frac{\mathbf{x}_i z_i}{\sigma_i^2}, \tag{17}$$

with $\mathbf{x}_i$ the vector $(x_{1,\,i}, \cdots, x_{J,\,i})^{\mathrm{T}}$ and σ_i^2 the model-variance of the ith unit. So, similar to the distinction between model mean and population mean, here the model regression coefficients $\boldsymbol{\beta}$ are distinguished from the population regression coefficients $\mathbf{B}$. Given these population regression coefficients $\mathbf{B}$, the model means $\mu(\mathbf{x}_i)$ are estimated by the means $m(\mathbf{x}_i)$, which are computed by:

$$m(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{B}. \tag{18}$$

If we have a probability sample from the population of interest, $\mathbf{B}$ is estimated by replacing the population totals in Eqn. 17 by their Horvitz-Thompson estimators:

$$\hat{\mathbf{B}} = \left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}}{\sigma_i^2\pi_i}\right)^{-1}\sum_{i=1}^{n}\frac{\mathbf{x}_i z_i}{\sigma_i^2\pi_i}, \tag{19}$$

with π_i the inclusion probability of the ith sampling location. Note that with unequal inclusion probabilities, the design-based estimates of the regression coefficients differ from the usual ordinary least squares estimates of the regression coefficients defined as model parameters. The values $\hat{B}_j$ are estimates of the *population parameters* B_j.

The mean values $m(\mathbf{x}_i)$ are now estimated by:

$$\hat{m}(\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{B}} . \tag{20}$$

Plugging this Eqn. 20 into Eqn. 14 leads to the generalized regression estimator:

$$\hat{\bar{z}}_{\text{regr}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{B}} + \frac{1}{N}\sum_{i=1}^{n}\frac{z_i - \mathbf{x}_i^{\mathrm{T}}\hat{\mathbf{B}}}{\pi_i} . \tag{21}$$

Särndal et al. (1992) worked out the general regression estimator for various superpopulation models, such as the common ratio model, the group mean model, the group ratio model and the multiple regression model. So, the superpopulation model serves as a vehicle to derive an efficient model-assisted estimator.

An important property of model-assisted estimators is that, if a poor working model is used (our assumptions about how our population is generated are incorrect), then for moderate sample sizes the results are still valid, that is, the empirical coverage of a model-assisted estimate of the confidence interval of the population mean still is approximately equal to the nominal coverage. This is because the mismatch of the superpopulation model and the applied model-assisted estimator results in a large design-variance of the estimated population mean. This is illustrated, for instance, with a simulation study by Brus (2000): the ratio estimator, which assumes an intercept of 0, applied to a population generated by a linear regression model with an intercept of 10, was approximately unbiased and the confidence intervals were valid, even for the smallest sample size of 10. In contrast, if in the model-based approach a poor working model is used, then the predictions and the prediction error variances still are model-unbiased, but for the sampled population we may have serious systematic error in the estimated population mean, and the variance of local predictions may be seriously over- or underestimated. For that reason, model-based inference is also referred to as *model-dependent* inference (Table 3), stressing that we fully rely on the model, and that the validity of the estimates and predictions depends on the quality of the model (Hansen, Madow, & Tepping, 1983)

An interesting application of the model-assisted approach is small domain estimation. Small domains are in our case small subareas. When the domains are numerous and are not well represented in the sample, estimators that only use the data from the domain to be estimated may lead to large sampling variances. Extensive literature exists about how the means of such small domains can be estimated (see Chauduri (1994) and Ghosh and Rao (1994) for reviews). The generalized regression estimator proposed by Hidiroglou and Särndal (1985) also uses sampling locations outside the domain. The intention is that, by doing so, the precision will increase. On the other hand, in general some design-bias will be introduced.

3.1.2 | Model-assisted estimators based on penalized spline regression model

The generalized regression estimator can also be written as a weighted average of the sample data. These weights can become quite extreme, especially when applied for estimating means of small areas. Various model-assisted estimators have been proposed that try to avoid these extreme weights, through smoothing, trimming or otherwise (see Breidt and Opsomer (2017) for an extensive list of references). One of the options is to use a penalized spline regression model as a working model (Breidt, Claeskens, & Opsomer, 2005):

$$\begin{aligned} z_i &= \beta_0 + \beta_1 x_i + , \cdots , \beta_q x_i^q + \sum_{k=1}^{K} \gamma_k (x_i - \kappa_k)_+^q + \epsilon_i \\ &= m(x_i; \boldsymbol{\beta}, \boldsymbol{\gamma}) + \epsilon_i \end{aligned}, \quad (22)$$

with q the degree of the spline, κ_k the knots, and $(x_i - \kappa_k)_+^q = (x_i - \kappa_k)^q$ if $(x_i - \kappa_k) > 0$ and 0 otherwise. To avoid unstable fits, the influence of the knots is limited by constraining the size of the coefficients $\gamma_k, \cdots, \gamma_K$. If all population units were observed the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ would be estimated by minimizing:

$$\sum_{i=1}^{N} (z_i - m(x_i; \boldsymbol{\beta}, \boldsymbol{\gamma}))^2 + \lambda \sum_{k=1}^{K} \gamma_k^2. \quad (23)$$

The population-fit of all the regression coefficients, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, is the penalized least squares estimator:

$$\mathbf{B} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{z}, \quad (24)$$

with $\mathbf{X}$ the matrix

$$\begin{bmatrix} 1 & x_1 & \cdots & x_1^q & (x_1 - \kappa_1)_+^q & \cdots & (x_1 - \kappa_K)_+^q \\ \vdots & & & & \vdots & & \\ 1 & x_N & \cdots & x_N^q & (x_N - \kappa_1)_+^q & \cdots & (x_N - \kappa_K)_+^q \end{bmatrix}$$

and $\boldsymbol{\Lambda}$ a diagonal matrix with $q + 1$ zeros on the diagonal, followed by K penalty constants λ. The sample-based estimator of $\mathbf{B}$ is:

$$\hat{\mathbf{B}} = \left(\mathbf{X}_s^{\mathrm{T}}\mathbf{W}\mathbf{X}_s + \boldsymbol{\Lambda}\right)^{-1} \mathbf{X}_s^{\mathrm{T}}\mathbf{W}\mathbf{z}_s, \quad (25)$$

with $\mathbf{X}_s$ the submatrix of $\mathbf{X}$ with the rows of $\mathbf{X}$ for the population units in the sample, $\mathbf{z}_s$ the sample data of the study variable, and $\mathbf{W}$ the matrix with the inverse of the inclusion probabilities of the sampling points. The above results are conditional on the number and locations of the knots. Opsomer, Claeskens, Renalli, Kauermann, and Breidt (2008) mention for a univariate spline (one covariate) one knot for every four or five observations, with a maximum of 35–50. The knots can best be spread out in covariate space (Ruppert, 2002).

If no covariates are available, we still can use the spatial coordinates in fitting a two-dimensional spatial spline. This is not entirely straightforward as we now have two 'covariates', the two spatial coordinates,

whereas in the spline regression model of Eqn. 22 we have only one. The question is how to construct a bivariate (spatial) spline. A straightforward solution is to extend the design matrix $\mathbf{X}$ with extra columns, so that it has in total $2K$ columns with truncated polynomials, K columns per covariate (spatial coordinate). This may lead to numerical instability. To reduce the number of columns, Ruppert, Wand, and Carroll (2003) proposed replacing the $2K$ columns with truncated polynomials with K columns with so-called pseudo-covariates that are a function of the Euclidian distance from the sampling points to the knots. Cicchitelli and Montanari (2012) applied this two-dimensional spatial version of the penalized spline regression model as a working model in the model-assisted difference estimator. They showed that the gain in efficiency can be impressive compared to the Horvitz-Thompson estimator.

As an illustration I simulated a population with a model mean of 25, and an exponential model without nugget, with a distance parameter of 20 distance units and a sill-variance of 100 (Figure 12a). The population mean equals 16.03. A stratified simple random sample is selected of 50 points, two points per stratum, with strata square blocks of equal size. The fitted spline shown in Figure 12b is a two-dimensional spatial penalized B-spline, fitted with R package SAP (Rodríguez-Álvarez, Lee, Kneib, Durbán, & Eilers, 2015). A penalized B-spline has several advantages over a penalized spline computed from truncated polynomials (Eqn. 22). For an explanation of penalized B-splines I refer to Eilers, Currie, and Durban (2006) and Eilers, Marx, and Durbán (2015). I repeated the stratified random sampling 200 times, fitted a two-dimensional penalized B-spline to the samples, and used the fitted values in the model-assisted difference estimator. The variance of the 200 estimates equals 0.619. The standard error of the Horvitz-Thompson estimator of the population mean equals 0.829. However, the average of the 200 variance estimates equals 0.346, showing that the true variance is strongly underestimated. Cicchitelli and Montanari (2012) explained this by overfitting of the sample data by the spline model, so that the residuals of the sampled units are smaller than those of the non-sampled units. More research on (approximately) unbiased estimation of the variance of the difference estimator exploiting two-dimensional spatial spline model predictions, is needed.

Opsomer et al. (2008) worked out a model-assisted estimator based on a spatial version of the penalized spline regression model for estimating the means of numerous small subareas. The spatial coordinates are used to construct the spline, following the approach of Ruppert, Wand, and Carrol (2003). Their working model is a linear mixed model with two random effect terms, one for the spline part of the model, and one for the small subareas. They illustrated their approach by estimating the mean acid binding capacity (ANC) of lakes in the north-eastern states of the USA. The population consists of 21,026 lakes; 334 of them were surveyed. The mean ANC of the lakes within 113 small areas was estimated. In 27 areas no observations were available. Besides the random effect for the small areas and a smooth spatial

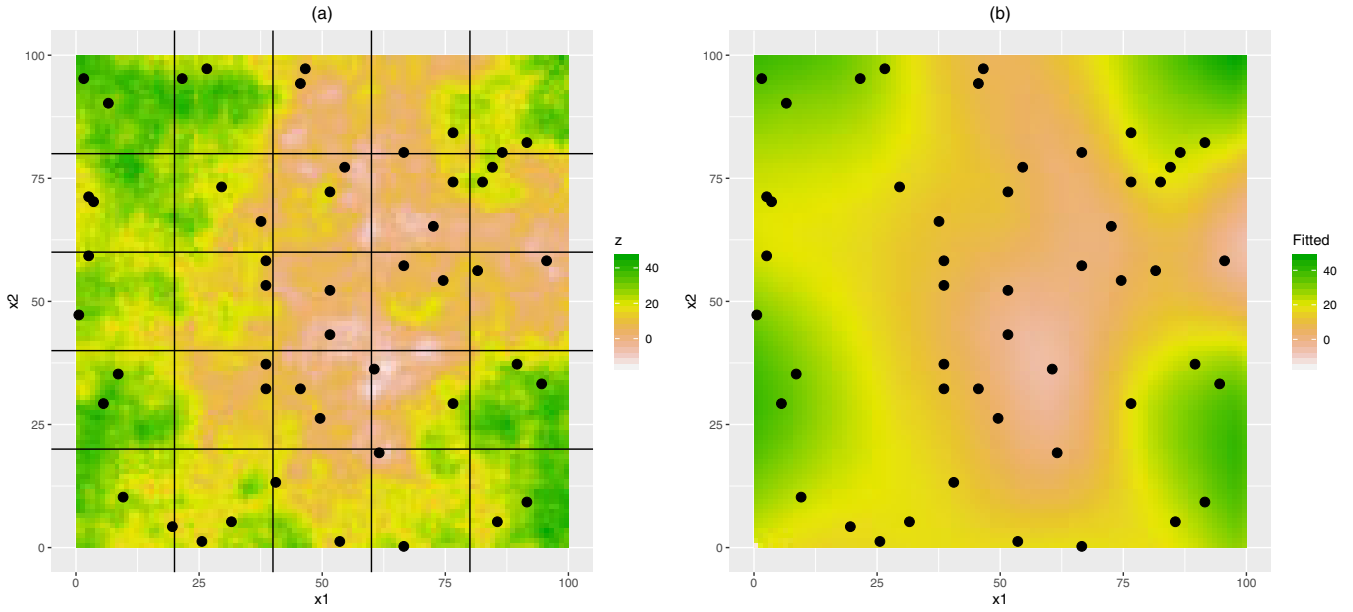

FIGURE 12 Simulated population with stratified simple random sample of 50 points (a) and fitted two-dimensional spatial penalized B-spline (b)

trend captured by the spline random effect, a linear elevation effect was incorporated in the model. Eighty knots were used, the locations of which were selected by a space-filling design.

3.1.3 | Model-assisted estimators based on statistical learning techniques

Breidt and Opsomer (2017) also review model-assisted estimators based on statistical learning techniques. Of special interest is the general approach proposed by Wu and Sitter (2001) for incorporating non-linear predictions in the model-assisted estimator. They show how non-linear predictions of the study variable, for instance obtained by a random forest model, can be used in the model-calibration estimator:

$$\hat{\bar{z}}_{\mathrm{MC}} = \hat{\bar{z}}_{\mathrm{HT}} + \hat{B}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{m}(\mathbf{x}_i) - \frac{1}{N}\sum_{j=1}^{n}\frac{\hat{m}(\mathbf{x}_j)}{\pi_j}\right), \tag{26}$$

with $\hat{B}$ a regression coefficient estimated by:

$$\hat{B} = \frac{\sum_{i=1}^{n} 1/\pi_i \hat{m}(\mathbf{x}_i) z_i}{\sum_{i=1}^{n} 1/\pi_i \hat{m}(\mathbf{x}_i)^2}. \tag{27}$$

The variance of the model-assisted calibration estimator equals:

$$V\left(\hat{\bar{z}}_{\mathrm{MC}}\right) = V\left(\hat{\bar{u}}_{\mathrm{HT}}\right), \tag{28}$$

with $\hat{\bar{u}}_{\mathrm{HT}}$ the Horvitz-Thompson estimator of the population mean of the residuals u. These residuals are equal to $u_i = z_i - B\, m(\mathbf{x}_i)$, with $m(\mathbf{x}_i)$ the fitted values based on all population units, and B the population fit of the regression coefficient:

$$B = \frac{\sum_{i=1}^{N} 1/\pi_i \hat{m}(\mathbf{x}_i) z_i}{\sum_{i=1}^{N} 1/\pi_i \hat{m}(\mathbf{x}_i)^2}. \tag{29}$$

An estimator of the variance is obtained by replacing the population fits $m(\mathbf{x}_i)$ by their sample estimates $\hat{m}(\mathbf{x}_i)$, and B by its estimator (Eqn. 27).

For simple random sampling with replacement from finite populations and simple random sampling from infinite populations the variance equals:

$$V\left(\hat{\bar{z}}_{\mathrm{MC}}\right) = \frac{S^2(u)}{n}, \tag{30}$$

with $S^2(u)$ the population variance of the residuals u.

Overfitting of the sample data by a random forest model, leading to underestimation of the true sampling variance of the estimated population mean (see Section 3.1.2) can be avoided by using the out-of-bag predictions for the units in the sample.

The alternative to the model-calibration estimatir is to plug the fitted values $\hat{m}(\mathbf{x}_i)$ into the difference estimator (Eqn. 14). For non-linear working models these two estimators are not the same. The model-calibration estimator has smaller variance than the model-assisted difference estimator (Wu & Sitter, 2001).

Viscarra-Rossel, Brus, Lobsey, Shi, and McLachlan (2016) used the Cubist model as a working model to estimate the total soil C stock in an area. A mobile multi-sensor platform was used to collect data of various covariates, which were used in Cubist to map soil organic C. The cubist predictions were subsequently used as a covariate in the calibration estimator (Eqn. 12 in Viscarra-Rossel et al., 2016).

3.2 | Modelling the sampling process in a model-based approach

As shown in Section 2.3, model-unbiasedness does not fully protect against selection bias, resulting in a systematic error in the model-based predictions, computed as an average of the prediction errors over all population units. Methods have been developed to reduce this systematic error, by accounting for the selection probabilities of the sampling locations. Diggle, Menezes, and Su (2010) proposed modelling the sampling locations along with the observations of the study variable in a joint model. Both the sampling locations and the observations are modelled as a function of an unobservable Gaussian process, referred to as the signal. More specifically, the point process for the sampling locations is modelled as an inhomogeneous Poisson process, in which the natural log of the intensity parameter of the Poisson distribution, that is, the parameter for the number of points per unit area, is a linear combination of the signal.

To estimate the model parameters, that is, the variogram parameters of the signal, the two coefficients for the intensity parameter of the Poisson process, and the measurement error variance, Diggle et al. (2010) approximated the likelihood function of the data by a Monte Carlo sample average, obtained by simulation of the signal conditional on the observations of the study variable. As noted by the authors, the convergence of the

estimation algorithm is slow and running time may become burdensome, needing a large number of Monte Carlo samples. Moreover, Dinsdale and Salibian-Barrera (2019) argued that the Monte Carlo estimate may not approximate the actual likelihood function, because the approximation is implicitly conditioned on the sampling locations. Dinsdale and Salibian-Barrera (2019) proposed approximating the likelihood function by differentiation of a Laplace approximation to the likelihood, instead of Monte Carlo approximation.

Pennino et al. (2018) proposed a Bayesian approach for fitting the parameters of a similar model. Their model consists of two linked models, one for the spatial point pattern and one for the observations (referred to as marks). The two models are linked through a shared Gaussian random field: both the model mean of the data and the intensity parameter of the point pattern model are a function of the same Gaussian random field. To account for differences in scale (variance) the Gaussian random field in the point pattern model is multiplied by a constant. The model parameters are estimated by integrated nested Laplace approximation (Rue, Martino, & Chopin, 2009), and by approximating the Gaussian random field through the partial differential equation approach (Lindgren, Rue, & Lindström, 2011).

4 | CONCLUSIONS

There is ongoing confusion about the design-based approach for spatial sample surveys, which has led to new publications with wrong formulas for the variance of an estimated population mean and the required sample size. The confusion is caused by ignorance of the fundamental difference between the design-based and model-based approaches due to the different sources of randomness that are accounted for in the two approaches. This difference in source of randomness means that basic statistical concepts, such as independence, variance and bias, have a different meaning in the two approaches. To avoid confusion, I recommend being more specific when using these terms, unless this is clear from the context, by adding the adjective design- or model-, for instance design-independent and model-variance. Also, now and then the population mean is confused with the model mean, and the model-variance with the population variance.

Both approaches are valid and have their strengths and weaknesses. Broadly speaking, the design-based approach is most appropriate for estimating the population mean or the means of several subpopulations, and its strength is the validity of the estimates. For mapping, the model-based approach is most appropriate and its strength is its efficiency. Various hybrid methods have been developed that try to combine the strengths of the two approaches. The use of a superpopulation model as a working model in the design-based approach has led to the model-assisted approach. If a working model is used that explains a reasonable part of the spatial variation of the soil property of interest, the efficiency is increased, whereas the validity of the estimates is maintained. The usefulness of a wide variety of models has already been shown for spatial sample survey, amongst others penalized spline regression models and non-linear models based on statistical learning techniques.

In the model-based approach models have been developed modelling the spatial variation of the data along with the point pattern of the sampling locations. These models may reduce the systematic error in model-based predictions due to preferential sampling.

ACKNOWLEDGEMENTS

I thank Jaap de Gruijter for his valuable comments on a draft version of this paper, and Martin Boer for his help with fitting the 2D penalized B-spline.

ORCID

Dick J. Brus https://orcid.org/0000-0003-2194-4783

REFERENCES

Baillargeon, S., & Rivest, L.-P. (2011). The construction of stratified designs in R with the package stratification. *Survey Methodology*, *37*, 53–65. http://www.statcan.gc.ca/pub/12-001-x/2011001/article/11447-eng.pdf

Barnes, R. J. (1988). Bounding the required sample size for geologic site characterization. *Mathematical Geology*, *20*, 477–490.

Breidt, F. J., Claeskens, G., & Opsomer, J. D. (2005). Model-assisted estimation for complex curves using penalised splines. *Biometrika*, *92*, 831–846.

Breidt, F. J., & Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, *32*, 190–205.

Brus, D. J. (2000). Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science*, *51*, 159–172.

Brus, D. J. (2015). Balanced sampling: A versatile sampling approach for statistical soil surveys. *Geoderma*, *253–254*, 111–121.

Brus, D. J., & de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, *80*, 1–59.

Brus, D. J., Spätjens, L. E. E. M., & de Gruijter, J. J. (1999). A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. *Geoderma*, *89*, 129–148.

Brus, D. J., Yang, L., & Zhu, A. X. (2019). Accounting for differences in costs among sampling locations in optimal stratification. *European Journal of Soil Science*, *70*, 200–212.

Chauduri, A. (1994). Small domain statistics: A review. *Statistica Neerlandica*, *48*, 215–236.

Cicchitelli, G., & Montanari, G. E. (2012). Model-assisted estimation of a spatial population mean. *International Statistical Review*, *80*, 111–126.

Clifford, P., Richardson, S., & Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics*, *45*, 123–134.

Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.

de Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., & Knotters, M. (2006). *Sampling for natural resource monitoring*. Berlin: Springer.

de Gruijter, J. J., Minasny, B., & McBratney, A. B. (2015). Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. *Journal of Survey Statistics and Methodology*, *3*, 19–42.

de Gruijter, J. J., & ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, *22*, 407–415.

Deville, J. C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, *91*, 893–912.

Diggle, P. J., Menezes, R., & Su, T. (2010). Geostatistical inference under preferential sampling. *Applied Statistics*, *59*, 191–232.

Dinsdale, D., & Salibian-Barrera, M. (2019). Methods for preferential sampling in geostatistics. *Journal of the Royal Statistical Society, Applied Statistis, Series C*, *68*, 181–198.

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.

Eilers, P. H. C., Currie, I. D., & Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, *50*, 61–76.

Eilers, P. H. C., Marx, B. D., & Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, *39*, 149–186. https://www.raco.cat/index.php/SORT/article/view/302258

Ghosh, M., & Rao, J. (1994). Small area estimation: an appraisel. *Statistical Science*, *9*, 55–93.

Grafström, A., Lundström, N. L. P., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, *68*, 514–520.

Griffith, D. A. (2005). Effective geographic sample size in the presence of spatial autocorrelation. *Annals of the Association of American Geographers*, *95*, 740–760.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability sampling inferences in sample-surveys. *Journal of the American Statistical Association*, *78*, 805–807.

Hidiroglou, M. A., & Särndal, C. E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology*, *11*, 65–77.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 423–498.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove: Duxbury Press.

Opsomer, J. D., Claeskens, G., Renalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*, 265–286.

Papritz, A., & Webster, R. (1995). Estimating temporal change in soil monitoring: I. statistical theory. *European Journal of Soil Science*, *46*, 1–12.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., & Conesa, D. (2018). Accounting for preferential sampling in species distribution models. *Ecology and Evolution*, *9*, 653–663.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. Boca Raton: CRC Press.

Ripley, B. D. (1981). *Spatial statistics*. New York: John Wiley & Sons.

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M., & Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: The SAP algorithm. *Statistics and Computing*, *25*, 941–957.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computations l Graphical Statistics*, *11*, 735–757.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.

Stevens, D. L., & Olson, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of American Statistical Association*, *99*, 262–278.

Viscarra-Rossel, R. A., Brus, D. J., Lobsey, C., Shi, Z., & McLachlan, G. (2016). Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. *Geoderma*, *265*, 152–163.

Wang, J., Haining, R., & Cao, Z. (2010). Sample surveying to estimate the mean of a heterogeneous surface: Reducing the error variance through zoning. *International Journal of Geographical Information Science*, *24*, 523–543.

Wang, J., Stein, A., Gao, B., & Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, *2*, 1–14.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists* (2nd ed.). Chichester: Wiley.

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*, 185–193.

How to cite this article: Brus DJ. Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *Eur J Soil Sci*. 2020;1–18. https://doi.org/10.1111/ejss.12988