Design-based and model-based sampling strategies for soil monitoring

Dick Brus[A]

[A]Soil Science Centre, Wageningen University and Research Centre, P.O. Box47, 6700 AA Wageningen, The Netherlands, Email dick.brus@wur.nl

Abstract
This paper explains the fundamental differences between the design-based and the model-based approach for sampling. In soil monitoring four combinations of these two approaches are possible, a fully design-based approach, a fully model-based approach, and two mixed, design-based and model-based approaches. The choice between these four approaches is crucial in designing a sampling scheme for monitoring. Another important choice is the pattern type of the observations in space-time, differing in how many and when sampling locations are revisited. Five basic types are described. Two case studies are then described, and the choices of the statistical approach and the pattern type are motivated.

Key Words
Probability sampling, trend monitoring, compliance monitoring, validity.

Two fundamentally different sampling approaches
In sampling for soil *survey* two fundamentally different approaches can be followed: a design-based or a model-based approach (Särndal *et al.* 1992; de Gruijter and ter Braak 1990). In a design-based approach sampling locations are selected by probability sampling, and the statistical inference (e.g. estimation of spatial mean) is based on the sampling design. In a model-based sampling approach there are no requirements on the method for selecting sampling locations, and typically are selected by purposive (targeted) sampling, for instance on a centred grid. In statistical inference a model for the spatial variation is introduced, e.g. an ordinary kriging model, assuming a constant (unknown) mean, or a universal kriging model in which the mean is modelled as a linear function of one or more predictors. Besides the deterministic part for the mean, a kriging model contains a stochastic part describing the variance and covariance of the residuals of the mean. Note that the source of randomness is different in the two approaches. In the design-based approach the selection of the sampling locations is random, whereas in the model-based approach randomness is introduced via the model of spatial variation. In the design-based approach no such model is used. This has important consequences for the interpretation of measures of uncertainty about estimates, e.g. the variance of the estimation (prediction) error.

To quantify our uncertainty about estimates (predictions) in both approaches a chance experiment is repeated many times (not in reality but in mind). However, as the source of randomness differs between the two approaches, this chance experiment also differs. In the design-based approach the chance experiment entails repeated selection of sampling locations with a random sampling design, whereas in the model-based approach a long series of values is generated at all locations in the area with a model, i.e., a series of `fields' (model-realizations) is simulated. Note that in the design-based approach only one `field' is considered, being the 'field' actually sampled, and in the model-based approach only one sample is considered, being the sample actually selected.

Choosing between the two approaches is one of the most important decisions in designing sampling schemes (de Gruijter *et al.* 2006). Brus and de Gruijter (1997) discuss pros and cons of both approaches, and give rules for choosing. Broadly speaking, a design-based approach is the best choice if interest is (parameters of) the spatial cumulative distribution function (SCDF) for the area as a whole or for a restricted number of subareas, and besides validity of the result really matters (validity more important than efficiency). A model-based approach is the best option if interest is in a map depicting the values of many small areas, e.g. pixels, and we want to predict these values as precise as possible (efficiency more important than validity). We try to increase the precision by postulating a model, which may invoke discussions on the validity of the result, as several assumptions in modelling are made.

Sampling for monitoring
In sampling for soil *monitoring* one dimension is added, the time dimension. Besides *where* to observe the

soil and at how many locations, we must decide on *when* to observe it, and how frequent. Similar to sampling locations, sampling times can be selected either by probability sampling or by non-probability sampling, the former enhancing design-based statistical inference, for instance of (parameters) of the temporal cumulative distribution function (TCDF), the latter asking for model-based inference. Having two options for spatial sampling and two options for sampling in time, four combinations for sampling in space *and* time are obtained:

1. D_SD_T: both sampling locations and sampling times are selected by probability sampling, and statistical inference is entirely design-based
2. D_SM_T: sampling locations are selected by probability sampling, but sampling times are not. Inference involves both design-based and model-based inference. The model used in model-based inference is a time-series model
3. M_SD_T: sampling locations are not selected by sampling, but sampling times are. Design-based and model-based inference. The model now is a spatial model
4. M_SM_T: neither sampling locations nor sampling times are selected by probability sampling, and a full space-time model is used in the inference

The choice between these four statistical approaches is crucial in designing sampling schemes for monitoring. Another important choice is the type of sampling pattern in space-time. Several basic types can be distinguished, differing in how many locations and when sampling locations are revisited (Figure 1). In a *static* pattern sampling locations are fixed (static), but the observations are not synchronized. In a *synchronous* pattern the observations are synchronized, i.e., all locations are observed in short periods of time (sampling rounds). However, the sampling locations observed differ between the sampling rounds, there are no revisits. In a *static-synchronous* pattern in all sampling rounds the same sampling locations are observed, i.e., all locations are always revisited. Two compromise patterns in between a synchronous and a static-synchronous pattern can be thought of. In a *rotational* pattern, part of the sampling locations of the previous sampling round is revisited, the other part is replaced by new locations (sampling with partial replacement). In a *serially alternating* pattern no locations are revisited until a given sampling round, after which all locations of the first round are revisited *et cetera*. Note that the method of selection of sampling locations or times (probability or non-probability) is not part of the definition of the pattern types. In figure 1 locations are selected randomly, but sampling times are not.

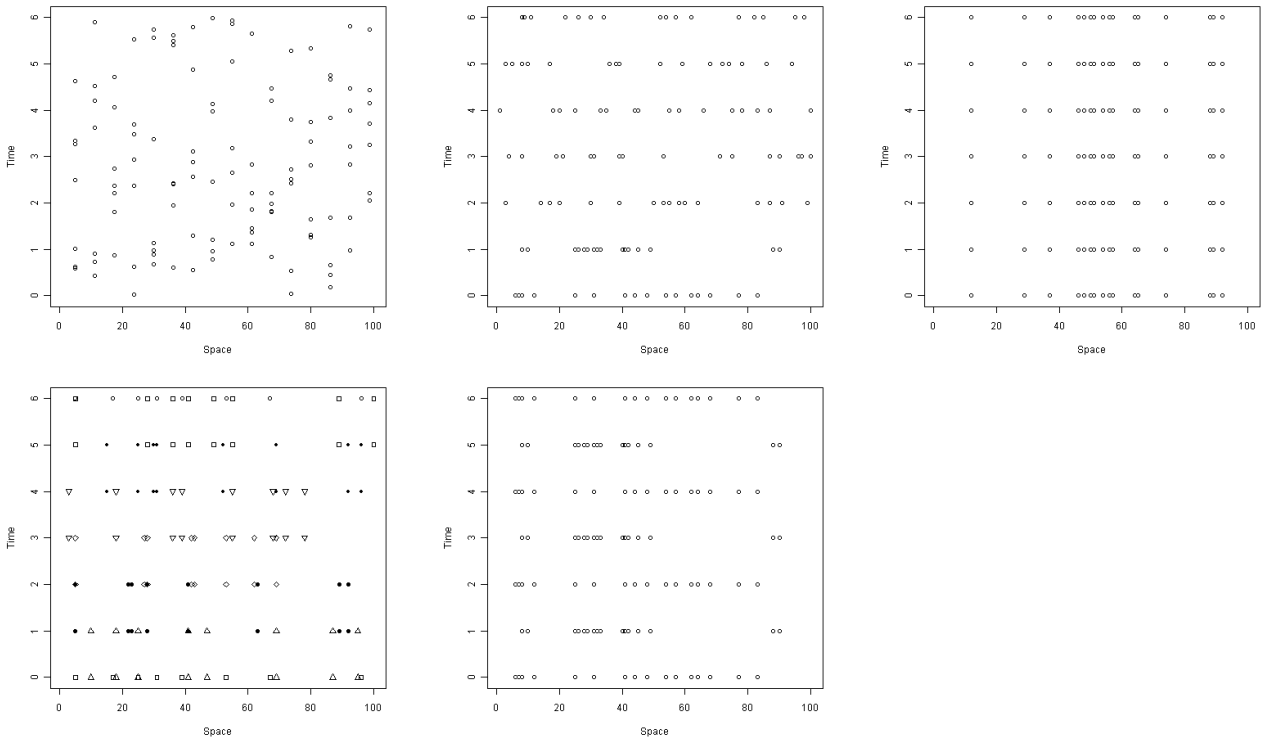

Figure 1. Basic types of sampling pattern in space-time. From top to bottom and from left to right: static, synchronous, static-synchronous, rotational (matching proportion 50%), and serially alternating (after de Gruijter *et al* (2006))

Case studies

I will illustrate now the above mentioned possibilities with two case studies, one on compliance monitoring of surface water quality, the second one on trend monitoring of indicators for soil eutrophication and acidification under forests. The choices with regard to statistical approach and pattern type that have been made will be motivated. In the first case study a fully design-based approach D_SD_T was chosen (Brus and Knotters 2008; Knotters and Brus 2010). The reason is that in compliance monitoring the quality of the result, being the conclusion whether the surface water quality complies with legal standards or not, must be beyond discussion. In other words, the validity of the result is very important. Moreover, interest was in a global target quantity, being the space-time mean concentration during a summer half-year. There was no need for spatial mapping of concentrations. As a pattern type a synchronous pattern was chosen, in which the sampling locations of a given round were selected *independently* from the locations of any other round. This independent synchronous sampling enables *design-unbiased* estimation of the sampling variance. For a static-synchronous pattern an unbiased estimator of the sampling variance does not exist. This is due to the two-fold alignment in space-time (Figure 1), i.e., the sampling locations of a given round are not selected independently from the locations of the other rounds, they even coincide with the locations of other rounds. This is entirely comparable to systematic random sampling in space (random grid sampling), for which an unbiased estimator of the sampling variance does not exist either. Both sampling locations and sampling times were selected by stratified simple random sampling. The stratification along both the spatial and the time axis improved the coverage of the space-time universe.

In trend-monitoring of soil eutrophication and acidification indicators a hybrid, design-based and model-based approach D_SM_T was chosen (Brus and de Gruijter 2010). So sampling locations were selected randomly, but sampling times were selected non-randomly, with the first sampling round at the start, the last one at the end of the monitoring period. The reason for selecting locations randomly is based on the chosen target quantity, being the temporal trend *of the spatial mean*. The available budget allowed for twenty locations per sampling round, which makes high-resolution mapping of temporal trends unfeasible. For estimating a temporal trend, random selection of times is suboptimal. For estimating a linear trend it would be optimal to do half of the total number of observations at the start, the other half at the end. However, this prevents us from checking for a non-linear trend. A rotational pattern type was chosen. This choice was somewhat arbitrary as we lacked knowledge of the relative efficiency of space-time pattern types for estimating the selected target quantity. However, we were reluctant to revisit in all sampling rounds twenty locations only, which would lead to a rather poor coverage of the space-time universe. To estimate the temporal trend of the spatial means, first the spatial means at the sampling rounds and their sampling variances and covariances were estimated by model-free, design-based inference. The estimated spatial means were then plotted against time. A linear trend was then fitted by generalized least squares. The covariance matrix used in GLS is the sum of the matrix with sampling variances and covariances of the estimated spatial mans and a matrix with model variances and covariances of the (errorless) spatial means. The hybrid approach enables quantification of the contribution to the total uncertainty about the trend of the sampling error in the estimated spatial means and of the model inadequacy error.

References

Brus DJ, de Gruijter JJ (1997) Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma* **80,** 1-59.

Brus DJ, Knotters M (2008) Sampling for compliance monitoring of surface water quality. A case study in a polder area. *Water Resources Research* **44,** W11410, doi:10.1029/2007WR006123.

Brus DJ, de Gruijter JJ (2010) A hybrid, design-based and model-based sampling approach for estimating temporal trends of spatial means. *Journal of Agricultural, Biological and Environmental Statistics*, under review

de Gruijter JJ, Brus DJ, Bierkens MFP, Knotters M (2006) 'Sampling for Natural Resource Monitoring'. (Springer: New York).

de Gruijter JJ, ter Braak CJF (1990) Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology* **22,** 407-415.

M Knotters, Brus DJ (2010) Estimating space-time mean concentrations of nutrients in surface-waters of variable depth. *Water Resources Research* (In press).

Särndal CE, Swensson B, Wretman J (1992) 'Model-assisted Survey Sampling'. (Springer: New York).