How to define, sample for and estimate the regional trend in soil monitoring?

From D.J. Brus
Soil Science Centre,
Wageningen University and Research Centre

1 Introduction

In soil monitoring we are often interested in whether the soil property of interest has been changed. Think for instance of changes in the soil carbon stock. With more than two sampling times we may be interested in the average change per time unit (for instance decade), which is equivalent to the *linear* trend. The average change per time unit generally will vary in space, some soil profiles respond quickly, others slowly. When we do not have enough budget for *mapping* the linear trend at point-locations, an alternative aim is to estimate the regional trend, defined as the linear trend of the spatial mean of the soil property of interest. In recent papers we have shown that this linear trend can be defined in different ways (Brus and de Gruijter, 2011, 2012). In this short paper I will elaborate on these definitions and illustrate sampling strategies for the trend with a simulated space–time field of soil organic matter (SOM) content (Figure 1)

2 Trend defined as population parameter

The linear trend can be defined as as a linear combination of the spatial means at the sampling times:

$$b = \frac{\sum_{j=1}^{r}(t_j - \bar{t})(\bar{z}_j - \bar{\bar{z}})}{\sum_{j=1}^{r}(t_j - \bar{t})^2} \tag{1}$$

with r the number of sampling times, $\bar{t}$ the mean of the sampling times, and $\bar{\bar{z}}$ the mean of the spatial means. You may recognize this as the Ordinary Least Squares (OLS) estimator of the slope of a linear model for $\bar{z}$ (dependent or response variable) and t as predictor. However, here the trend is not a model parameter, but a population parameter. The population or universe of interest consists of a finite set of (infinite or finite) spatial populations, $\mathcal{U} = \{\mathcal{S}_1, \mathcal{S}_2 \cdots \mathcal{S}_r\}$, with $\mathcal{S}_1$ the spatial population at sampling time t_1, *et cetera*. This universe is a subset only of the

How to define, sample for and estimate the regional trend in soil monitoring?

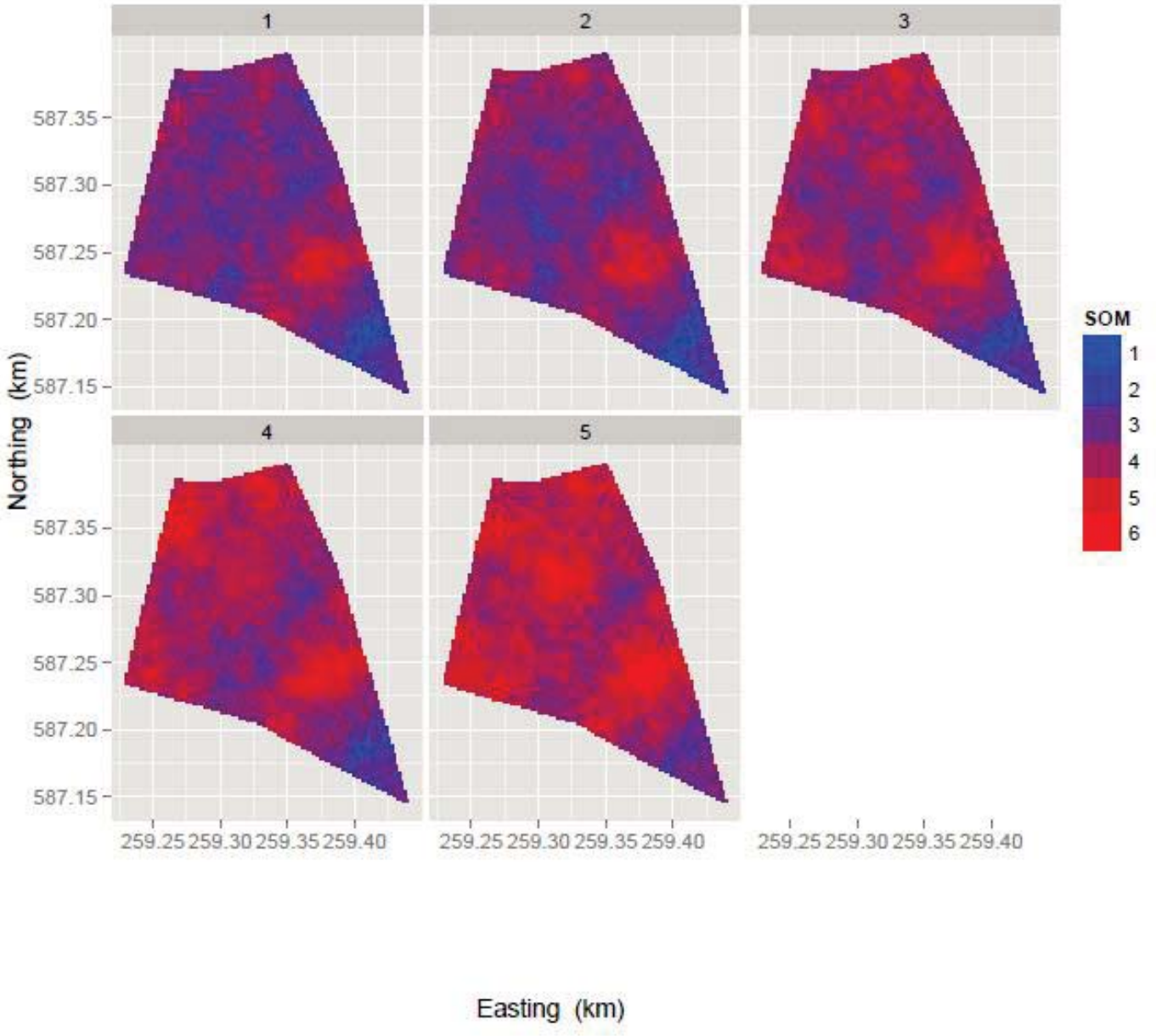

Figure 1: Simulated space–time field of soil organic matter content. The five panels show the spatial fields at the sampling times. The sampling interval is constant (e.g. 10 years).

How to define, sample for and estimate the regional trend in soil monitoring?

$\mathcal{U} = \mathcal{S} \times \mathcal{T}$ with $\mathcal{T}$ the temporal universe (ter Braak et al., 2008). I will not go into sampling approaches for this definition of the trend.

Parameter b as defined in Eq. 2 can also be seen as the slope parameter that is obtained when the response variable is known for all population units (exhaustive fit). Here the population 'units' are not sampling units (points) but populations themselves, viz. the spatial populations at the r sampling times. The response variable is the *spatial mean* of SOM. When the spatial means are known for all population units, i.e. at at all sampling times, then parameter b is also known without error, see hereafter.

Eq. 2 can be rewritten as a linear combination of the spatial means at the sampling times:

$$b = \frac{\sum_{j=1}^{r}(t_j - \bar{t})\bar{z}_j}{\sum_{j=1}^{r}(t_j - \bar{t})^2} = \sum_{j=1}^{r} w_j \bar{z}_j \tag{2}$$

with the weights w_j equal to

$$w_j = \frac{t_j - \bar{t}}{\sum_{j=1}^{r}(t_j - \bar{t})^2} \tag{3}$$

This shows that the trend can be estimated via estimation of the spatial means at the sampling times, and as a consequence a design-based sampling approach is recommendable. I will elaborate now on estimation for space–time designs with no or complete overlap (static-synchronous, independent synchronous, serially alternating) and for space-time designs with partial overlap (supplemented panel, rotating panel).

2.1 Space–time designs with no or complete overlap

With space–time designs in which the spatial samples at the sampling times $t_1 \cdots t_r$ have no overlap, i.e. no locations are revisited, or complete overlap, i.e. all locations are revisited, the spatial mean at a given time is estimated on the basis of the measurements at that time only, using the well-known design-based estimators. Given these estimated means the linear trend can be estimated as a linear combination of the estimated means

$$\hat{b} = \sum_{j=1}^{r} w_j \hat{\bar{z}}_j = \mathbf{w}' \, \hat{\mathbf{z}} \tag{4}$$

2.2 Space–time designs with partial overlap

For space–time designs with partial overlap such as the supplemented and the rotating panel, the precision of the estimated mean at a given sampling time can be increased by using the measurements at the other times as covariates. This can be

achieved by Generalized Least Squares (GLS) estimation of the spatial means. First panel-specific estimates of the spatial means are computed, referred to as 'elementary estimates'. A panel is a group of locations observed at the same set of sampling times. These elementary estimates are then combined into one estimate of the mean per time t_j by

$$\hat{\mathbf{z}}_{\mathrm{GLS}} = (\mathbf{X}'\widehat{\mathbf{C}}_{\mathrm{e}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{C}}_{\mathrm{e}}^{-1}\hat{\mathbf{z}}_{\mathrm{e}} \tag{5}$$

with $\hat{\mathbf{z}}_{\mathrm{e}}$ the vector of elementary estimates of the spatial means, $\mathbf{X}$ the design matrix with 0's and 1's, and $\widehat{\mathbf{C}}_{\mathrm{e}}$ the estimated covariance matrix of the elementary estimates. Hopefully you recognize this equation from your statistics courses on regression analysis as the GLS-estimator of the regression coefficients. In linear regression analysis we have observations on a target variable and one or more predictors, covariates. In ordinary linear regression it is assumed that the observations are independent. Correlation between the observations can be accounted for by estimating the variance-covariance matrix of the observations, and using this matrix in GLS fitting of the linear model. Here the observations of the target variable are the elementary estimates of the spatial means at $t_1 \cdots t_r$. The predictors are indicators for the sampling times. There are as many predictors as there are sampling times.

Once the means are estimated by GLS, the trend can be estimated as a linear combination of these estimated means:

$$\hat{b}_{\mathrm{GLS}} = \mathbf{w}'\hat{\mathbf{z}}_{\mathrm{GLS}} \tag{6}$$

with $\mathbf{w}$ as before (Eq. 3). With small spatial sample sizes the estimated sampling covariance matrix $\widehat{\mathbf{C}}_{\mathrm{e}}$ can be poorly defined, leading to extreme values for the estimated trend. In such cases I recommend to estimate the trend with Eq. 4.

2.3 Effect of number of sampling locations on variance of estimated trend

The sampling variance of the estimated trend can be reduced by increasing the number of sampling times and the number of sampling locations per time. Besides, there is a clear effect of the type of space–time design (Fig. 2) and of the spatial design. Fig. 3 shows the standard error of the estimated trend as a function of the number of sampling locations per time, for a static-synchronous space–time design and simple random sampling in space. If the entire study area would be sampled all five times, the standard error would be 0. There is no uncertainty left about the trend. Figure 3 (subfigure in the middle) shows the true spatial means at the five times plotted against the sampling time and the estimated linear trend. As can be seen the true spatial means are not located precisely on the fitted line. In regression analysis we would say that there is a residual variance. As a consequence in regression analysis the variance of the estimated regression coefficients (intercept and slope) is not 0 but a positive value.

How to define, sample for and estimate the regional trend in soil monitoring?

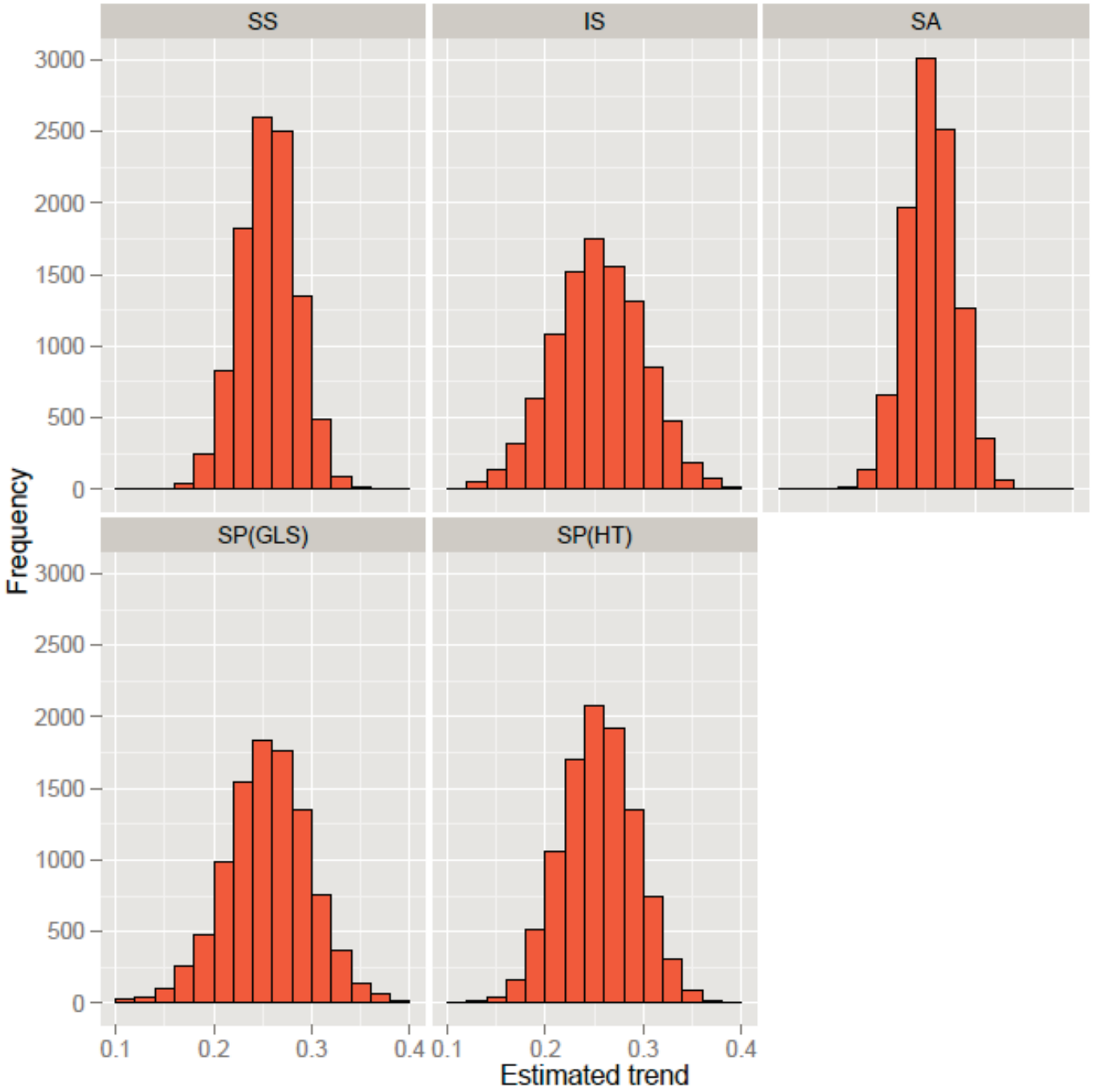

Figure 2: Histograms of 10,000 repeated estimates of the trend of the mean defined as population parameter, for static-synchronous (SS), independent-synchronous (IS), serially alternating (SA) and supplemented panel (SP) sampling, five sampling times and 20 locations per time selected by simple random sampling (sampled from the space–time field of Figure 1). In supplemented panel sampling 10 locations are revisited. For SP the trend is estimated both by Eq. 4 (SP(HT)) and by Eq. 6 (SP(GLS)). Note the long tails of the sampling distribution of the estimated trend with SP(GLS), caused by the poorly defined covariance matrix. The serially alternating design had the smallest sampling variance of the estimated trend

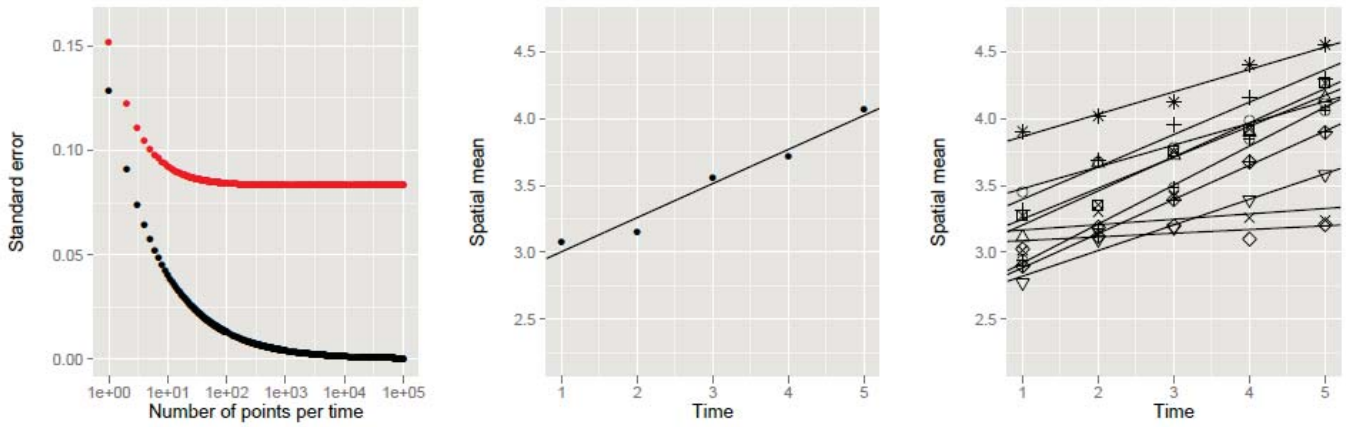

Figure 3: Left: standard error of the estimated trend, defined as a population parameter (black) or as a model parameter (red), as a function of the number of points per time. Sampling design: static-synchronous with simple random sampling in space. Middle: true spatial means of simulated space–time field (Fig. 1), plotted against the sampling time, and the linear trend of the spatial means. Right: true spatial means and linear trend fitted by OLS for 10 realizations of the space–time model used in simulating Fig. 1

3 Trend defined as model parameter

Fitting the straight line of Fig. 3 by OLS with standard statistical software results in an estimated trend of 0.255 which is equal to the estimated trend defined as a population parameter. However, the standard deviation of the estimated trend equals 0.027, which is small, but definitely larger than 0. The reason that in standard regression analysis the variance is not 0 is that the true spatial means are considered as realizations of random variables. In OLS fitting of the simple linear model the spatial means at the sampling times are assumed to be identically and independently distributed with expectation $\beta_1 + \beta_2 \cdot t$ and constant variance (the variance of the residuals). The coefficients β_1 and β_2 are model parameters, the intercept and the slope, respectively. The parameter β_2 describes the average change of the spatial mean per time unit, the linear temporal trend. This is the target parameter to be estimated. So, contrary to the previous section a time-series model is introduced for the spatial means

$$\bar{Z}(t_j) = \beta_1 + \beta_2 \cdot t_j + \eta(t_j) \quad j = 1 \cdots r \tag{7}$$

where $\eta(t_j)$ is the model residual (model error) of the spatial mean at time t_j. The spatial mean at time t_j is now in capital, indicating that it is a random variable. With the trend defined as a model parameter the sampled space–time field of Fig. 1 is treated as just one realization of a stochastic space–time process. I simulated 10 of these space–time fields, computed for each simulated space–time field the true

spatial means at the five times, and fitted the model by OLS. The result is presented in Fig. 3 (subfigure at the right). The fitted trend clearly varies between the model-realizations. The variation is even much larger than expected from the estimated variance of the trend as obtained with OLS fitting (standard deviation 0.027). This can be explained by the correlation of the spatial means. In OLS it is assumed that these spatial means are uncorrelated (identically independently distributed, iid), however this assumption is clearly violated by the space–time model used in simulating the space–time fields. The spatial means are correlated in time, amplifying the variance of the trend between model realizations. Fig. 3 (subfigure on the left) shows that the standard error of the estimated trend, defined as a model parameter, with exhaustive spatial sampling is about 0.083.

In practice the spatial means are unknown, and must be estimated from a sample. When these spatial means are estimated from probability samples and design-based estimators, then the space–time sampling approach becomes a hybrid, design- and model-based approach. To explain this approach I will first consider the simple situation where we have only one estimate of the mean per time, and then proceed with the situation with more than one elementary estimate per time, as obtained with space–time designs with partial overlap.

3.1 Space–time designs with no or complete overlap

In the hybrid approach it is assumed that the spatial means can be described by a linear mixed model

$$\mathbf{Z} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta} \,, \tag{8}$$

with $\mathbf{Z}$ the r-vector with true spatial means at the sampling times, $\mathbf{D}$ the $r \times p$ design-matrix, $\boldsymbol{\beta}$ the p-vector with regression coefficients and $\boldsymbol{\eta}$ the r-vector with model errors. This matrix equation is equivalent to the model of Eq. 7 for a design-matrix $\mathbf{D}$ with the first column a vector of ones and the second column a vector with the sampling times. The model errors $\boldsymbol{\eta}$ have zero mean and an $r \times r$ covariance matrix $\mathbf{C}_\xi$. This is the matrix with the variances and covariances of the spatial means between realisations of the space–time model. In practice the spatial means are unknown, and in the hybrid approach these means are estimated from spatial probability samples. With no or complete overlap these spatial means are estimated by design-based estimators. The sampling introduces an additional error component in the model:

$$\widehat{\mathbf{Z}} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon} \,, \tag{9}$$

with $\boldsymbol{\epsilon}$ the r-vector with sampling errors. The sampling errors have zero mean and an $r \times r$ covariance matrix $\mathbf{C}_p$, the sampling covariance matrix of the estimated spatial means that we have seen many times before. The model errors and sampling errors are independent, as they originate from independent stochastic processes. The

overall covariance matrix of the estimated spatial means equals

$$\mathbf{C}_{\xi p} = \mathbf{C}_{\xi} + \mathbf{C}_{p} \, . \tag{10}$$

With known covariance matrix $\mathbf{C}_{\xi p}$, the regression coefficients can be estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\hat{\mathbf{Z}} \tag{11}$$

3.2 Space–time designs with partial overlap

Model 9 is reformulated so that multiple estimates of the spatial mean at a given time are accounted for:

$$\widehat{\mathbf{Z}}_{\mathrm{e}} = \mathbf{D}_{\mathrm{e}}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\eta} + \boldsymbol{\epsilon}_{\mathrm{e}} \, . \tag{12}$$

Design matrix $\mathbf{D}_{\mathrm{e}}$ now has dimension $E \times p$ with E the total number of elementary estimates. $\mathbf{X}$ is a random-effect design matrix (dimension $E \times r$) with zeroes and ones selecting the appropriate element of $\boldsymbol{\eta}$. Vector $\boldsymbol{\eta}$ is as before, but vector $\boldsymbol{\epsilon}_{\mathrm{e}}$ now has length E as we have multiple sampling errors per sampling time, one per elementary estimate. The overall covariance matrix of the estimated spatial means equals

$$\mathbf{C}_{\xi p} = \mathbf{X}\mathbf{C}_{\xi}\mathbf{X}' + \mathbf{C}_{ep} \, . \tag{13}$$

with covariance matrix $\mathbf{C}_{\xi}$ as before (dimension $r \times r$) and $\mathbf{C}_{ep}$ the sampling covariance matrix of the elementary estimates (dimension $E \times E$). With known covariance matrix $\mathbf{C}_{\xi p}$, the regression coefficients can be estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{D}_{\mathrm{e}}'\mathbf{C}_{\xi p}^{-1}\mathbf{D}_{\mathrm{e}})^{-1}\mathbf{D}_{\mathrm{e}}'\mathbf{C}_{\xi p}^{-1}\hat{\mathbf{Z}}_{\mathrm{e}} \tag{14}$$

With small spatial sample sizes the estimated covariance matrix can be not positive definite or poorly defined, leading to missing values or extreme estimates. In this case a simple alternative is to estimate the spatial means at the sampling times by the design-based estimators, as well as as their sampling variances and covariances, and then proceed as in the previous section for samples with no or complete overlap.

4 Which definition?

The question remains what definition can best be chosen. I think the definition is at least partly determined by the aim of the monitoring project. If the aim is to describe the trend during the monitoring period, then a definition in terms of a population parameter is more appropriate than as a model parameter. A definition of the trend as a model parameter comes into scope if we want to use the results for forecasting, i.e. predicting the status in the future. If we use the estimated trend of the mean defined as a population parameter and its standard error for this, then this may lead

to too optimistic estimates of the precision. Clearly, for forecasting the structure of the trend is extremely important. In the case study on SOM a linear trend might not be very realistic when forecasting over long terms. It is more likely that the trend is asymptotically towards a maximum (or minimum in case of a negative trend), which can be modelled, for instance, by an exponential decay (in increasing or decreasing form). In this case the aim would be to estimate the parameters of this exponential model.

Another factor that may help in choosing a definition can be the feasibility of the statistical sampling approach. The definition of the trend has implications for the statistical sampling approach. When defined as a model parameter, a hybrid approach is needed. This sampling approach requires the calibration of a time-series model for the spatial means. which can be difficult. The more sampling times, the more information is obtained on the model. With a few sampling times only, the building of the model can become unfeasible. Strong assumptions are then needed, for instance on stationarity of the spatial mean and on the covariogram model. Besides, the model parameter estimates may become very unreliable. The quality of the estimates, especially the variance of the estimated regional trend, depends on the quality of these assumptions and estimates. With a few sampling times only, we might prefer a model-free sampling approach. Judging a hybrid sampling approach as unfeasible entails that we must abandon the trend defined as a model parameter, and embrace the trend defined as a population parameter as in Eq. 2 as the space–time parameter to be estimated.

References

Brus, D. J. and de Gruijter, J. J. (2011). Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data. *Geoderma*, 164:172–180.

Brus, D. J. and de Gruijter, J. J. (2012). A hybrid design-based and model-based sampling approach to estimate the temporal trend of spatial means. *Geoderma*, 173-174:241–248.

ter Braak, C. J. F., Brus, D. J., and Pebesma, E. J. (2008). Comparing sampling patterns for kriging the spatial mean temporal trend. *Journal of Agricultural, Biological and Environmental Statistics*, 13:159–176.