Approximating the sampling variance of means estimated from systematic random sample data of the french soil monitoring network

Nicolas N. Saby, D. J. Brus, Hakima Boukir, Vera Laetitia Mulder

APPROXIMATING THE SAMPLING VARIANCE OF MEANS ESTIMATED FROM SYSTEMATIC RANDOM SAMPLE DATA OF THE FRENCH SOIL MONITORING NETWORK

INTRODUCTION

The sites of the French soil monitoring network are selected by systematic random sampling (SY). It consists of a 16 x 16 km grid, leading to a total of about 2200 sites. These sites are sampled every 10 to 15 years. SY leads to good spatial coverage, i.e. the sites are uniformly spread over France, enhancing the precision of design-based estimates of spatial means and totals. Besides, SY is a suitable sampling design for spatial mapping e.g. by kriging. SY therefore is a flexible sampling design: SY samples can be used both for design-based estimation of means and totals, and for mapping.

With SY the sample average is an unbiased estimate of the mean. **No unbiased estimator exists for the sampling variance of this estimator** This paper investigates **5 approximations** of the sampling variance using a *simulation study* and *a case study* in France on several soil properties.

MATERIALS

The study was performed in mainland France, excluding Corsica, and covers approximately 542.0*10^3 km^2. First, the map units of the 1:1 000 000 soil map of Europe were aggregated on the basis of information on soil parent material (level 1) within these map units. We selected **4** aggregated units (see figure). We selected also the **watershed** of the Loire river.

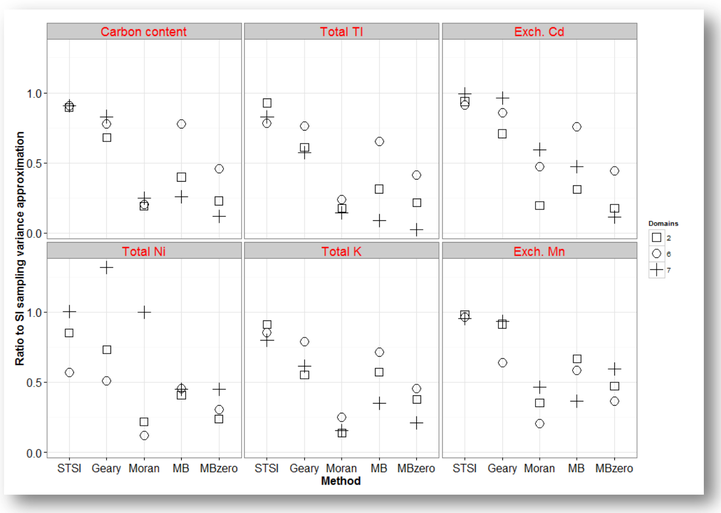
In the *simulation study* the true sampling variance of the estimated mean of NDVI for these three units was determined by repeated SY sampling. The approximated sampling variances were compared with this true sampling variance. In the *case study* the data of the first campaign of the French Soil Monitoring Network (FSMN) were used for approximating the sampling variances of the estimated means of several topsoil properties.

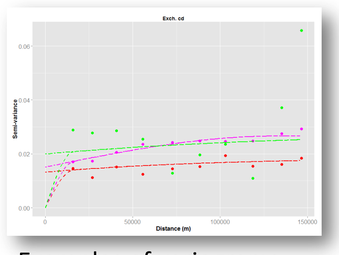
METHOD

Variance approximation	Name	Expression
Simple Random Sampling	SI	$\tilde{V}_{SI}(\hat{\bar{z}}) = \frac{s^2}{n}$
Stratified simple random sampling (with two neighboring cells as strata)	STSI	$\tilde{V}_{ST}(\hat{\bar{z}}) = \sum_{h=1}^{L} \left(\frac{n_h}{n}\right)^2 \frac{s_h^2}{n_h}$
Model-based prediction	MB, MBZero*	$\tilde{V}_{MB}(\hat{\bar{z}}) = \bar{\gamma} - \mathrm{E}_p[\bar{\gamma}_C]$
Simple random sampling corrected for serial correlation among sample unit with Geary's G index	Geary	$\tilde{V}_G(\hat{\bar{z}}) = \tilde{V}_{SI}(\hat{\bar{z}}).c$
Simple random sampling corrected for serial correlation among sample unit with Moran's *I* index	Moran	$\tilde{V}_I(\hat{\bar{z}}) = \tilde{V}_{SI}(\hat{\bar{z}}).r$

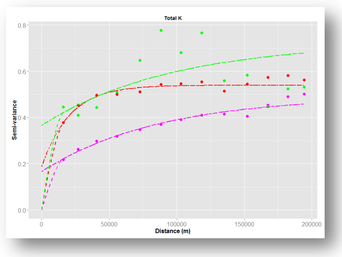
* Model-based prediction requires a variogram. A square grid is not ideal for estimating a variogram, because no pairs of points at short distance are available, which are crucial for estimating the nugget parameter. For that reason we computed two model-based approximations, one with the variogram as fitted (MB), and one with a variogram in which the fitted nugget parameter is replaced by a spherical model with a sill equal to the fitted nugget parameter and a range equal to the grid-spacing (16 km). (**MBZero)**

RESULTS FOR THE CASE STUDY





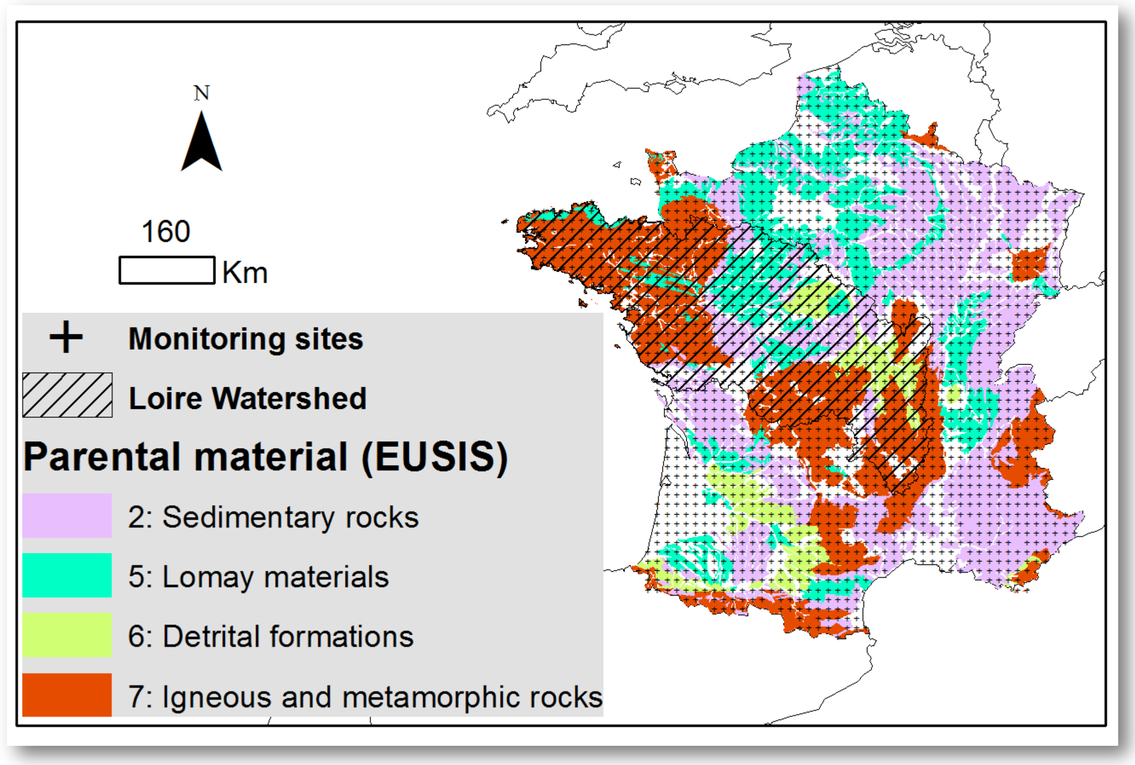
Exemples of variograms



DISCUSSION & CONCLUSIONS

- ✓ SI always strongly overestimated the sampling variance, whereas Moran always underestimated the variance; both should therefore not be used
- ✓ STSI and Geary were on average the best approximation methods
- ✓ MBzero approximated the variance better than MB for two out of the three units
- ✓ The variogram is central in the MB method

Study area, monitoring sites and domains

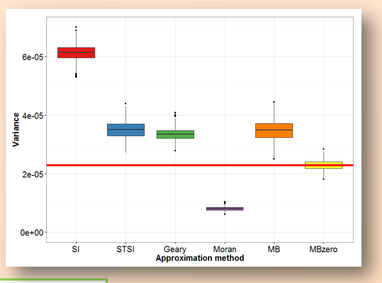


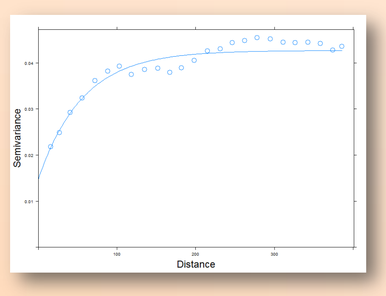
SIMULATION STUDY

A map of NDVI as obtained by MODIS was used as reality. The map was sampled one million times by SY, using the 16 km grid-spacing of FSMN. For each SY sample from a parent material unit the mean of NDVI was estimated. The variance of the ? estimated means served as the benchmark value (horizontal line in figure). Each SY sample was used to approximate the sampling variance (box plots)

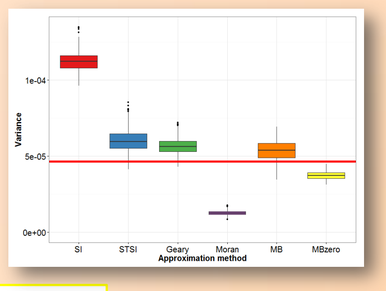
Experimental and fitted variograms for the different domains computed with **one** replicate of the simulated SY samples

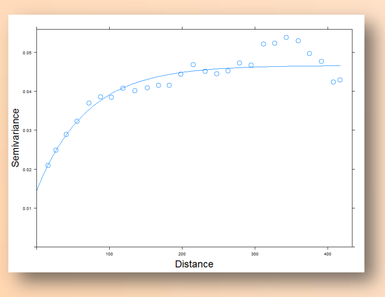
Domain 2



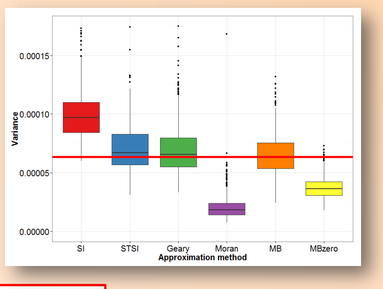


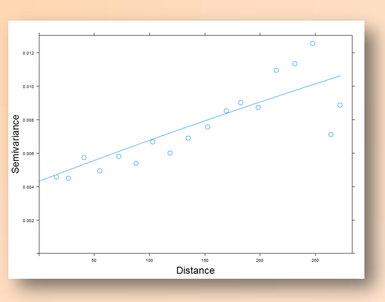
Domain 5



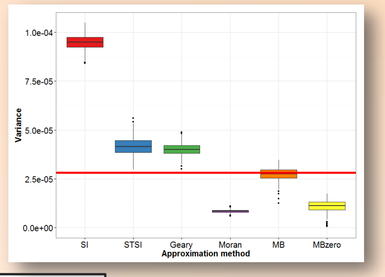


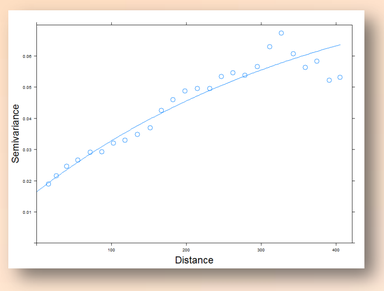
Domain 6



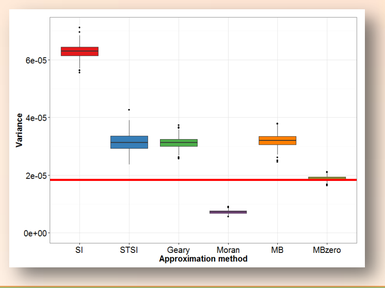


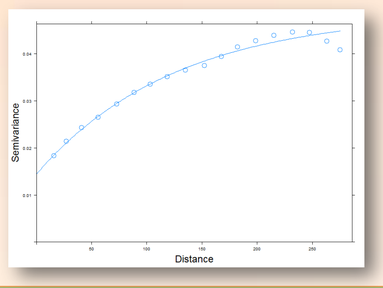
Domain 7





Watershed





N.P.A. Saby[a], D.J. Brus[b], H Boukir [a], V.L. Mulder[a]
a INRA, US1106, Unité Infosol, 45075 Orléans, France.
b Soil Science Centre, Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands