Geoderma 279 (2016) 77–86

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Approximating the variance of estimated means for systematic random sampling, illustrated with data of the French Soil Monitoring Network

D.J. Brus [a,*], N.P.A. Saby [b]

[a] Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, the Netherlands
[b] INRA, US1106 Unité Infosol, F-45000 Orléans, France

ARTICLE INFO

Article history:
Received 3 February 2016
Received in revised form 6 May 2016
Accepted 22 May 2016
Available online 9 June 2016

Keywords:
design-based inference
Moran's *I*Geary's spatial autocorrelation index
variogram
carbon stock

ABSTRACT

In France like in many other countries, the soil is monitored at the locations of a regular, square grid thus forming a systematic sample (SY). This sampling design leads to good spatial coverage, enhancing the precision of design-based estimates of spatial means and totals. Design-based estimation of the mean or total from SY samples is straightforward. However, an unbiased estimator of the sampling variance of the estimated mean or total does not exist. This paper compares five variance approximations, being the simple random (SI), stratified simple random (STSI), Geary's spatial autocorrelation *C* index (Geary), Moran's *I* index (Moran), and the model-based (MB) approximation in a simulation study and a real-world case study. In a simulation study the model distribution of the conditional bias (conditioned on a simulated reality) of the variance approximations is estimated for various variograms and two sample sizes. In the case study the data of the first campaign of the French Soil Monitoring Network are used to estimate the spatial means of six soil variables (C, Tl, Cd, Ni, K, Mn) for aggregated soil map units of France, and to approximate their sampling variances. The bias in the approximated variances is explored with MODIS-NDVI data. With variograms with no or a small relative nugget variance approximation STSI and MB are the best choices. In situations with large relative nugget STSI is to be preferred over MB as MB then may somewhat underestimate the variance. Moran and SI should be avoided as approximation methods, as they seriously underestimate (Moran) and overestimate (SI) the variance in many cases. The approximated standard error of total soil organic carbon stock in France as obtained with MB was 0.0335 Pg, which was small compared to the estimated stock of 3.580 Pg.

1. Introduction

Recently scientists and politicians have become aware that soils are not an unlimited resource and are an important resource supporting life on Earth (McBratney, 2016). Indeed soil resources provide many important ecosystem goods and services. However, they are at risk from a variety of threats operating over a broad range of scales. Although rates of soil degradation are often slow and only detectable over long timescales, they are often irreversible. Therefore, monitoring soil quality is essential in order to detect adverse changes in their status at an early stage. To implement soil protection policies and identify where soil protection measures are required it is necessary to have accurate information on how soil properties vary. Thus there is a need for high intensity national-scale soil surveys (Morvan et al., 2008). The data gathered from these surveys must be analysed by reliable statistical methods which are appropriate throughout the region of interest to ensure that the quantified uncertainty associated with the results are not distorted by statistical artefacts.

Like in many other countries 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 5 to 25 years. SY leads to good spatial coverage, i.e. the sites are uniformly spread over France. This is profitable both for mapping soil properties and for estimating spatial means or totals of these properties (think, for instance, of soil carbon stocks). When sampling locations are selected by probability sampling, spatial means and totals can be estimated by design-based inference (Brus and de Gruijter, 1997). Many soil properties show spatial contiguity to some extent. As a consequence, two locations that are close to each other contain less information about the mean then two locations far away from each other. This intuitively explains why random sampling designs leading to good spatial coverage, such as systematic random sampling (SY), provide more precise estimated means than sampling designs leading to spatial clusters of sampling units such as in simple random sampling (Quenouille, 1949; Cochran, 1977).

Systematic spatial sampling on a regular grid is also a suitable sampling design for mapping soil properties by (model-based) spatial interpolation, e.g. by kriging. When the origin of the grid is randomly selected the resulting systematic random sample can thus

* Corresponding author.
E-mail address: dick.brus@wur.nl (D.J. Brus).

http://dx.doi.org/10.1016/j.geoderma.2016.05.016
0016-7061/

be used both for design-based estimation of means or totals and for mapping. This sampling design therefore is a flexible and attractive sampling design for statistical soil surveys.

Design-based estimation of means or totals from SY samples is straightforward: the sample mean is an unbiased estimator of the population mean. However, this is not the case for the sampling variance of the estimated mean or total, i.e. the variance of the estimated mean (total) over repeated systematic random sampling. This sampling variance quantifies our uncertainty about the estimated mean, and is crucial in decision making. An unbiased estimator of this sampling variance does not exist unless two or more systematic random samples are selected independently from each other (Lohr, 1999). Based on the work of Wolter (1984, 1985), D'Orazio (2003) proposed several approximate variance estimators for two-dimensional spatial populations. Domburg et al., 1994 showed how the variance of the mean estimated by systematic random sampling can be predicted from a variogram. Subsituting the mean semivariances in this predictor by sample estimates leads to another variance approximation. In this paper five variance approximations are compared in a simulation study and a real-world case study. In the simulation study 1000 Gaussian fields are simulated with various variograms. With simulated realities the bias in the variance approximations can be determined. The aim of the simulation study is to find out whether the relative performance of the variance approximations is related to the variogram used in simulation. In the case study the data of the first campaign of the French Soil Monitoring Network are used to estimate the spatial means of six soil variables (C, Tl, Cd, Ni, K, Mn) for five aggregated map units of the soil map of France, and to approximate their sampling variances. The data are also used to estimate the total soil organic carbon stock in the whole territory of France and to approximate its sampling variance. The bias in these variances cannot be determined as the true variances are unknown. We therefore used MODIS-NDVI as a surrogate variable. This variable is exhaustively known so that the true sampling variance can be determined experimentally by repeated selection of SY samples. This experimentally derived variance can then be used to estimate the bias in the approximated variances of the estimated means of NDVI.

2. Theory

With systematic random sampling the sample average

$$\bar{z}_s = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{1}$$

is an unbiased estimator of the mean of a soil property z across the study region. In this equation n is the sample size (number of gridpoints), and z_i is the observation at the i^{th} gridpoint. More specific, the estimator is *design*-unbiased, which means that over repeated sampling with the systematic random design the expectation of this estimator of the mean equals the true spatial mean:

$$\mathrm{E_p}[z_s] = z \tag{2}$$

where $\mathrm{E_p}[\cdot]$ is the statistical expectation over repeated sampling under sampling design p (in this paper SY), and z is the spatial mean (population mean) of z. The sampling variance of the estimated mean (in this case the sample average, Eq. 1) is defined as

$$\mathrm{V_p}\left[\bar{z}_s\right] = \mathrm{E_p}\left[\left(\bar{z}_s - \mathrm{E_p}\left[\bar{z}_s\right]\right)^2\right] \tag{3}$$

For unbiased estimators of the mean (Eq. 2) this variance becomes

$$\mathrm{V_p}\left[\bar{z}_s\right] = \mathrm{E_p}\left[\left(\bar{z}_s - \bar{z}\right)^2\right] \tag{4}$$

In general SY leads to relatively accurate estimates of the spatial mean and total. This is because sampling locations are well-spread throughout the study area, so that redundant information due to spatial clustering of sampling locations is avoided. A drawback is that no unbiased estimator of the sampling variance exists. The reason is that we have selected only one cluster of sampling locations, so that there is no replication. The sampling variance can only be approximated. In this research several variance approximations are compared.

A first approximation is to treat the SY sample as a simple random sample (SI). The variance is then approximated by

$$\tilde{V}_{\mathrm{SI}} = \frac{s^2}{n} \tag{5}$$

with s^2 the variance of the observations in the sample (referred to as the *sample* variance):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z}_s)^2 \tag{6}$$

This approximation overestimates the sampling variance for populations showing positive autocorrelation, especially with small grid-spacings (Cochran, 1977).

D'Orazio (2003) proposed a variance approximation in which strata are formed by collapsing adjacent gridcells either vertically or horizontally. The two observations in collapsed gridcells are treated as independent observations in a stratum, and the sampling variance of the estimated mean is approximated by the variance estimator for stratified simple random sampling. Our second approximation is an adaptation of this, avoiding the choice between collapsing gridcells either vertically or horizontally. In our approximation the gridpoints are clustered into $L = n/2$ equally sized clusters by k-means, using the spatial coordinates of gridpoints as clustering variables (Fig. 1). When n is an odd number, $n/2$ is rounded downward, so that there is one cluster with three gridpoints (all other clusters have two gridpoints). The k-means clustering of the gridpoints into clusters of equal size can be done with *R*-package spcosa (Walvoort et al., 2010). The two (three) gridpoints of a cluster are treated as the sampling points of a stratum selected by simple random sampling. The sampling variance is then approximated by:

$$\tilde{V}_{\mathrm{STSI}} = \sum_{h=1}^{L}\left(\frac{n_h}{n}\right)^2 \frac{S_h^2}{n_h} \tag{7}$$

with n_h the number of gridpoints in stratum (cluster) h, and s_h^2 the variance of the two (three) measurements of the variable of interest in cluster h.

In our third approximation, proposed by D'Orazio (2003), the sampling variance is approximated by multiplying the SI approximation $\tilde{V}_{\mathrm{SI}}$ by Geary's spatial autocorrelation index C:

$$\tilde{V}_{\mathrm{Geary}} = C \times \tilde{V}_{\mathrm{SI}} \tag{8}$$

with C the estimate of Geary's spatial autocorrelation index obtained from a 2-D systematic sample (Cliff and Ord, 1981):

$$C = \frac{n-1}{2S_0}\frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_{i,j}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - z)^2} \tag{9}$$

with $S_0 = \sum_{i=1}^{n}\sum_{i\neq j}^{n} w_{i,j}$ and $w_{i,j} = 0$ if i and j are not neighbours and $w_{i,j} = 1/n_i$ if gridpoint i has n_i neighbours (referred to as row-normalized weights). We also tried binary weights $w_{i,j} = 1$ if i and j are neighbours, 0 else. We used $\sqrt{2}$ times the gridspacing as an upper bound to define the neighbourhood, so that the maximum number of neighbours