Journal of Statistical Planning and Inference 181 (2017) 52–61

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Model-based variance estimation in non-measurable spatial designs

Roberto Benedetti [a], Giuseppe Espa [b], Emanuele Taufer [b,*]

[a] Department of Economic Studies, University of Chieti-Pescara, Italy
[b] Department of Economics and Management, University of Trento, Italy

ARTICLE INFO

Article history:
Received 16 February 2016
Received in revised form 5 August 2016
Accepted 7 September 2016
Available online 8 October 2016

Keywords:
Spatial survey
Two-dimensional systematic sampling
Maximal stratification
Variogram
Gaussian random field

ABSTRACT

Two-dimensional systematic sampling and maximal stratification are frequently used in spatial surveys, because of their ease of implementation and design efficiency. An important drawback of these designs, however, is that no direct estimator of the design variance is available. In this paper estimation of the sampling variance of a total in a model-based context is considered.

The estimation strategy is based on the use of the sample variogram which can be either a non-parametric or a parametric one. Consistency of the estimators is discussed; simulations and an application to real data show the good performance of the proposed procedure in practice.

1. Introduction

In agricultural and environmental surveys statistical units are often defined using purely spatial criteria, i.e. units are defined using geographical coordinates; for details see Benedetti et al. (2015). Also, many National Statistical Institutes are increasingly geo-referencing their sampling frames by adding information regarding the exact position of each record.

An inherent and fully recognized feature of spatial data is that they are dependent, as expressed in Tobler's first law (Tobler, 1970). As a consequence, certain sampling schemes for spatial units and estimators can be defined by introducing a suitable model for spatial dependence within a model-based or model-assisted framework.

In this paper we will discuss and implement a model-based estimator of the variance for some spatial sampling designs; in particular we will concentrate on two-dimensional systematic sampling and one-per-stratum (or maximal stratification) sampling which are quite common for surveys where sampling units are spatially referenced. They are relatively simple to plan and implement; provide unbiased estimators of totals and, selecting samples that are well-spread over the study region, can even yield lower variability in design-based estimators (Cochran (1977, sec. 5.7 and p. 208); Fewster (2011)). This property is mainly justified by the literature on spatially balanced samples, according to which, for both empirical and theoretical reasons selecting samples that are spatially well distributed implies a gain in efficiency, particularly when we are dealing with populations positively autocorrelated or that follow a spatial trend (Stevens and Olsen, 2004; Grafström and Tillé, 2013). However the distinguishing characteristics of these designs are that the second order probabilities are equal to zero at least for close units that belong to the same stratum or that are within the step used in systematic sampling. This is a condition that brings us in the field of non-measurable designs and implies the impossibility to use a design-based estimator of the variance.

* Corresponding author.
E-mail addresses: benedett@unich.it (R. Benedetti), giuseppe.espa@unitn.it (G. Espa), emanuele.taufer@unitn.it (E. Taufer).

http://dx.doi.org/10.1016/j.jspi.2016.09.003
0378-3758/

Recall that a probability sampling design is measurable if all the inclusion probabilities of the first and second order are strictly positive. The positivity of the inclusion probabilities of the first order is a sufficient condition for an unbiased estimator of a total to exist (Fuller, 2009, p. 8). The condition of positivity of the inclusion probability of the second order, instead, makes it possible to calculate an unbiased (or approximately unbiased) estimator of the sample variance. Such design-based variance can be used to build design-based confidence intervals. For all the details see Särndal et al. (1992, sect. 2.4 and sect. 14.3) and Benedetti et al. (2015, p. 115).

Solutions to the problem of variance estimation in non-measurable designs (as defined above) discussed in the literature and used in practice can be divided into three broad groups: (i) ignoring the problem, i.e. using variance estimators derived from simple random sampling; (ii) post-stratification, i.e. aggregating strata or adjacent samples from systematic designs and using stratified variance estimators; (iii) modeling the process producing the finite population and exploiting this information to estimate the variance.

There seems to be increasing interest in the literature in using explicit model-based solutions to the problem of variance estimation in non-measurable designs even if one is primarily interested in design-based inference: general reference texts are Wolter (2007, Ch. 8) and Fuller (2009, sec. 5.3) and specific contributions for spatial data are those of Opsomer et al. (2012) and Bartolucci and Montanari (2006) which rely on linear models based on auxiliary variables, Fewster (2011) which applies a multinomial model to strip sampling and transect sampling and D'Orazio (2003) which applies corrections based on Moran's and Geary's spatial auto-correlation statistics to simple random sampling and post-stratification derived estimators of variance.

This paper is strictly connected to this stream of research, where a design-based inference for the mean or the total of a population is coupled with a model-based estimation of the variance. No auxiliary variables are involved, however, we will assume that there is a random field underlying the population units.

In principle the method proposed can be applied on any design (as shown by Proposition 1) and we expect that, as our simulations show, the gain in efficiency is greater the stronger the structure of dependence on the underlying field.

On the other hand the method is computationally intensive and we believe its practical relevance be at its highest in non-measurable designs as in the case of systematic sampling and stratified sampling with one unit per stratum for spatial data. For other cases, where unbiased estimators of the variance exist, these might be preferred alternatives in practice.

In this paper a full discussion of the maximal stratification case is presented while we analyze the performance of our estimator in two-dimensional systematic sampling by means of simulations.

To see things in another way, one could say that kriging techniques (see, e.g. Cressie, 1993) are exploited for estimating the variance. In this context it is worth mentioning Goovaerts (1997) and Wang et al. (2009, 2013) and the references therein which discuss using kriging in the context of mean estimation.

We would like to point out that stratification with more than one unit per stratum is not considered here, being a measurable design for which a design-unbiased variance estimator exists. In this direction one can consult, e.g., the recent contributions of Wang et al. (2016, 2012, 2010).

For an up-to-date and full discussion of the designs discussed here and their relevant applications in fields such as natural resource surveys, forestry inventories and soil sampling for precision agriculture see Benedetti et al. (2010, 2015), Gregoire and Valentine (2007), and Tan (2005).

In Section 2 the estimators are defined and discussed; in Section 3, using simulated data, comparisons with other estimators of the variance using either parametric and non-parametric forms of the variogram are provided and an application to the celebrated Mercer and Hall data is presented. Proofs of the results are in Appendix.

2. Estimators of the expected variance

2.1. Notation and assumptions

Let $\{Y_i, i \in T\}$ denote a random field, where T is an index set. In a general setting $T = \mathbb{Z}^2$ represents a 2-dimensional lattice, while for $T = \mathbb{R}^2$ one has a continuous random field. T can also represent a collection of spatial entities such as territorial economic or administrative units. This last setting is the one which interests most here as the case where there is a, possibly very large, finite population U of size N; in this case let $T = T_N \subset \mathbb{Z}^2$ with $|T_N| = N$, i.e. $|T|$ indicates the cardinality of T. The set T_N of territorial units can be thought to be embedded in some general stationary field $\{Y_i, i \in \mathbb{R}^2\}$.

Let $\bar{Y}_N = N^{-1} \sum_{i=1}^{N} Y_i$ be the mean of U and let T_n, $|T_n| = n, n < N$, denote a sample set of observations from T_N collected according to some sampling strategy. The primary object of investigation is a model-based estimation of the variance of a design-based, say $\bar{Y}_d$, estimator of $\bar{Y}_N$, where the suffix d indicates the sampling design. For example, in the case of a systematic design, $\bar{Y}_d = \bar{Y}_{sy}$, the simple mean of the systematic sample; in the case of a stratified sampling design $\bar{Y}_d = \bar{Y}_{st}$, a weighted mean of the strata means, see Cochran (1977) for further details.

Estimation of the *expected design variance* $\mathrm{E}[\mathrm{Var}(\hat{Y}_d)]$ in a model based context is considered, i.e. when the finite population is regarded as a random realization from a super-population model. In our case we will assume that the geo-referenced Y's satisfy:

Assumption 1. $\{Y_i, i \in T\}$ is a stationary random field with mean $\mathrm{E}(Y_t) = \mu$, covariance $\mathrm{Cov}(Y_i, Y_j) = C(i-j)$ and $\mathrm{E}(Y^4) < \infty$.

Note that we do not require the field to be isotropic as the covariance rests on the difference $(i-j)$ between locations only, nor we require that T_N represents a regular lattice although in Section 4, for simplicity, simulations based on regular grids are considered. We are going to investigate the behavior of estimators and parameters of the finite population U as T_n and T_N grow large. For this we will place the restriction that $\|i-j\| \geq \delta > 0, \forall i, j = 1, 2, \ldots, N$ where $\|\cdot\|$ indicates the Euclidean norm. This assures that the observed field increases in extent as N increases. We are not interested here in the case where a sample may become increasingly dense in some bounded region.

Furthermore, we ask that the covariances be absolutely summable, i.e. we set:

Assumption 2. For the field of Assumption 1 it holds that $\lim_{N\to\infty} \sum_{i,j} |\mathrm{Cov}(i-j)| < \infty$.

The above assumption essentially excludes random fields with long memory. Examples of random fields satisfying Assumptions 1 and 2 are the so-called spherical model used in geo-statistics (see, e.g. Mardia and Marshall, 1984, Matheron, 1971, Journel and Huijbregts, 1978), and the isotropic covariance model discussed by Whittle (1954). See also Leonenko and Taufer (2013) for models on a lattice with covariance functions satisfying Assumptions 1 and 2.

In the paper we will extensively use the variogram $E(Y_i - Y_j)^2 = 2\gamma(i-j)$, see, e.g. Cressie (1993) for further details. In practice we will exploit the information given by the correlation between units and use distance as an auxiliary information.

A capital letter will be used to indicate the unit values either in the sample and the finite population. When needed, sample and population are distinguished by the extended notation $\{Y_i, i \in T_n\}$ and $\{Y_i, i \in T_N\}$ respectively.

2.2. Sampling with unequal probabilities of selection

We begin with a simple random sampling scheme with unequal probabilities of selection: given the form of the variance of the celebrated Horvitz–Thompson estimator, it provides a natural justification for the use of the variogram. Defining with $\hat{Y}_{HT}$ the Horvitz–Thompson estimator of a population total, we have:

$$\mathrm{Var}(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \tag{1}$$

where $\pi_i = P(Y_i, i \in T_N)$ and $\pi_{ij} = P(Y_i, Y_j, (i,j) \in T_N), i, j = 1, 2, \ldots, N, i \neq j$ are respectively the first and second order inclusion probabilities.

In order to exploit the spatial location of units and construct the variogram, define $N(d)$ as the set of pairs of observations with spatial coordinates i and j such that $|i-j| = d$; more formally, $N(d) = \{(i,j) : |i-j| = d; i, j \in T_N\}$ and $|N(d)|$ its cardinality. With some abuse of notation we simply indicate that $d = 1, \ldots, D$ with D indicating the total number of distinct differences $|i-j|$. In practical applications an approximate distance d is used, implemented with a certain tolerance.

The following proposition links the expected variance of the HT estimator to the variogram and will introduce our estimation strategy; the proof is in Appendix.

Proposition 1. *Let $\{Y_i, i \in T_N\}$ be a stationary random field satisfying Assumption 1 and define*

$$g_1(\pi, d) = \sum_{|i-j|\in N(d)} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i\pi_j}, \qquad g_2(\pi) = \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i\pi_j}. \tag{2}$$

Then

$$\mathrm{E}[\mathrm{Var}(\hat{Y}_{HT})] = 2\sum_{d=1}^{D} g_1(\pi, d)\gamma(d) + \sigma^2 g_2(\pi). \tag{3}$$

Formula (3) can be seen as a general formula for our estimation strategy. Note that the quantities $g_1(\pi, d)$ and $g_2(\pi)$ are known as information on the distance and the inclusion probabilities are known in advance for the whole population. It follows that by substituting in (3) consistent estimators of the variogram γ and the variance σ^2 one gets a consistent estimator for the expected variance.

In the case where all units have the same probability of being selected, i.e. $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$, $i, j = 1, 2, \ldots, N$, the anticipated variance can be reduced to

$$\mathrm{E}[\mathrm{Var}(\hat{Y}_{HT})] = 2\frac{N}{n}\frac{(N-n)}{(N-1)}\sum_{d=1}^{D}|N(d)|\gamma(d). \tag{4}$$

This essentially corresponds to the approach suggested by Fuller (2009, sec. 5.3) where he also discusses the application to the case of designs with one unit per stratum.

In maximal stratification the first order inclusion probabilities are typically constant, while $\pi_{ij} = \pi_i\pi_j$ if units i and j are in different strata and $\pi_{ij} = 0$ if the two units are in the same stratum. In the case of kth step systematic sampling the second order inclusion probabilities are $\pi_{ij} = 1$ if units i and j, are at some kth step distance, and 0 otherwise.

2.3. Stratified sampling with one unit-per-stratum design and systematic sampling

Consider the case of a stratified sample with one unit per stratum. Let $\bar{Y}_{st}$ denote the sample mean of a stratified sample and $h = 1, \ldots, H$ the strata. In the case of a one-per-stratum design, letting $W_h = N_h/N$, we have the classical result

$$\mathrm{Var}(\bar{Y}_{st}) = \sum_{h=1}^{H}(1 - N_h^{-1})W_h^2 S_h^2. \tag{5}$$

Again, one can link the expected variance to the variogram by noting that:

$$S_h^2 = \frac{1}{(N_h - 1)}\sum_{i=1}^{N_h}(Y_i - \bar{Y}_h)^2 = \frac{1}{2N_h(N_h - 1)}\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}(Y_i - Y_j)^2. \tag{6}$$

From (6) under the super-population model and using the relation $\mathrm{E}(Y_i - Y_j)^2 = 2\gamma(i - j)$,

$$\mathrm{E}(S_h^2) = \frac{1}{N_h(N_h - 1)}\sum_{i=1}^{N_h}\sum_{j=1}^{N_h}\gamma(i - j). \tag{7}$$

In stratum h, let $N_h(d) = \{(i, j) : |i - j| = d,\ i, j = 1, 2, \ldots, N_h\}$ and suppose, in stratum h, $d = 0, 1, 2, \ldots, D$; then:

$$\mathrm{E}(S_h^2) = \frac{1}{N_h(N_h - 1)}\left[N_h\gamma(0) + \sum_{d=1}^{D}|N_h(d)|\gamma(d)\right]. \tag{8}$$

Using all sample data, a parametric or semi-parametric model for $\gamma(d)$ can be estimated and substituted in (8) to obtain an estimated variance for $\bar{Y}_{st}$ in a one per stratum design.

With the help of Lemmas 1 and 2 in Appendix, we can establish a consistency result for the model-based estimators. The proof is given in Appendix.

Proposition 2. *Let $\hat{\gamma}$ denote a consistent estimator of the variogram γ. Then, under Assumptions 1 and 2, for $H = o(N)$, as $N \to \infty$,*

$$\sum_{h=1}^{H}(1 - N_h^{-1})W_h^2\hat{S}_h^2, \quad \textit{with } \hat{S}_h^2 = \frac{1}{N_h(N_h - 1)}\left[N_h\hat{\gamma}(0) + \sum_{d=1}^{D}|N_h(d)|\hat{\gamma}(d)\right], \tag{9}$$

is a consistent estimator of $\mathrm{Var}(\bar{Y}_{st})$ *as* $N \to \infty$.

Note that in Proposition 2, the term S_h^2 is substituted by a general variogram approximation, not depending on h. In practice the two dimensional systematic sampling can be treated analogously where the expected variance can be estimated by substituting a variogram approximation to all differences in the sampled population.

An *R* code to compute the estimators is available from the authors upon request.

3. Comparisons of estimators by a Monte Carlo study

This section presents the results of some Monte Carlo experiments by which the performance of estimator (9) and that of some alternative estimators are compared. Comparisons include populations with different intensities of spatial dependence and three different variogram estimators.

In the first subsection details on the simulation design such as the generated populations, the variogram estimators used and the alternative estimators used as a benchmark are provided. In the second one the results of the simulation will be presented and discussed.

3.1. Simulation design

For our comparison four different populations are considered: the first consists of real data, while the remaining are simulated ones.

The real population considered is based on the data collected by Mercer and Hall (1911, Tab. 5) in 1910 summer by the *Rothamsted Experiment Station* of Harpenden, Hertfordshire, England. This data set is well known and has been used in several papers in spatial analysis such as Whittle (1954), Patankar (1954), Besag (1974), Ripley (1981) and Cressie (1993). Mercer and Hall collected data on the weight in lbs of the yield of grain and straw on a field divided in 500 approximately regular cells (or plots). The dimension of each plot was approximately 3.2×2.5 m over a one-acre uniform area. Then one plot was approximately 1/500 of an acre. In the paper we consider only the data on the wheat yield. The dataset is then a regular grid (or raster) of 20 (in direction South–North) $\times$ 25 (in direction West–East) = 500 cells.

The three remaining populations are composed of simulated data on regular grids of $20 \times 20 = 400$ cells. The choice of this dimension is due to the possibility of choosing two different sample sizes in two dimensional systematic sampling and maximal stratification sampling.

The three simulated populations are the following:

(i) A quadratic trend (QT) of the form $Y_i = (i_1 - 10)^2 + (i_2 - 10)^2 + 3 + \varepsilon$ with $\varepsilon \sim N(0, 4)$ where the pixel i in the grid has coordinates (i_1, i_2). The random variables ε in each cell are independent of each other.
(ii) A Gaussian random field ($GRF1$) $\mu = 5, \sigma^2 = 1$ and exponential correlation function $\rho(u) = \exp\{-u/\phi\}$ with $\phi = 3$.
(iii) A Gaussian random field ($GRF2$) $\mu = 5, \sigma^2 = 1$ and Gaussian correlation function $\rho(u) = \exp\{-(u/\phi)^2\}$ with $\phi = 3$.

The first simulated population represents a spatial trend while in the other two there is presence of autocorrelation: the first with an exponential variogram and the second with a Gaussian variogram. Both can be useful to see what happens to the estimators if the variogram used is not the correct one (e.g. using an exponential variogram in computing the estimators while the population is characterized by a Gaussian one).

For further details on Gaussian random fields, see Diggle and Ribeiro (2007, Ch. 3). For the four populations above, in order to have an idea of the strength of spatial dependence we have computed Moran's index either under the normality assumption and under randomization, using queen's and rook's neighborhoods obtaining quite similar results in the various cases. The value of the index under the hypothesis of normality and using a rook neighborhood is 0.3073 for MH data, 0.9534 for QT data, 0.7389 for GRF1 and 0.8873 for GRF2.

In order to compare the performance of our and alternative estimators, the real and simulated populations are divided into n domains (or strata), i.e. the $N = R \times C$ regular grid of units of the population is divided in non-overlapping blocks of $k = k_R \times k_C$ cells. In this way n strata, each formed by a regular grid of size $n_R \times n_C$ with $n_R = R/k_R$ and $n_C = C/k_C$ (both integers), are obtained.

In the actual simulation runs we have the following values:

(i) for the Mercer and Hall (MH) data, $k_R = k_C = 5$. Hence the initial $N = 500 = 20 \times 25$ units have been organized in 25 strata composed of 20 units (pixels in Fig. 1);
(ii) for the three simulated populations, for the $N = 400 = 20 \times 20$ units, two distinct scenarios are considered: $k_R = k_C = 5$ and $k_R = k_C = 4$ which respectively yield two regular grids of $n = 16$ and $n = 20$ strata. In a such a way the aggregation problem of adjacent cells is limited (for a discussion see Ripley, 1981, pp. 108–109) and a control for increasing sample size is introduced.

With self-explaining acronyms, in the output tables we will indicate the different simulate populations and sizes as QT_{16}, QT_{25}, $GRF1_{16}$, $GRF1_{25}$, $GRF2_{16}$, $GRF2_{25}$.

As far as the sampling procedures are concerned, in the case of maximal stratification the selection procedure has been repeated 1000 times and based on this we construct the empirical distribution of the estimators. In the case of two-dimensional systematic sampling, given the number possible samples is limited to stratum size (k), we simply selected all possible samples.

The performance of the proposed estimation strategy is compared either with estimators which explicitly consider the spatial nature of the problem or estimators which ignore the problem. We do not consider estimators based on auxiliary variables as the estimators proposed here do not and this situation is quite common in agricultural trials.

(i) The classical variance estimator of the HT total, denoted with $\hat{V}_{SRS}(\hat{Y})$ (see Cochran, 1977, p. 261), which just ignores the problem and treat the systematic and stratified samples as simple random samples (SRS).
(ii) The estimators proposed by D'Orazio (2003) which imply a correction of $\hat{V}_{SRS}(\hat{Y})$ by using either Geary (c) or Moran (I) spatial auto-correlation indexes (see, e.g., Cliff and Ord, 1981, Ch. 1 and 3 or Ripley, 1981, sec. 5.4); namely $\hat{V}_{SRS}(\hat{Y}) \cdot c$ and $\hat{V}_{SRS}(\hat{Y}) \cdot I$.

Finally, as far as variogram estimators are concerned, we consider three different estimation strategies:

(i) a moment based variogram estimator, i.e. for $\{Y_i, i \in T_n\}$,

$$\hat{\gamma}(d) = \frac{1}{2N(d)} \sum_{(i,j) \in N(d)} (Y_i - Y_j)^2; \tag{10}$$

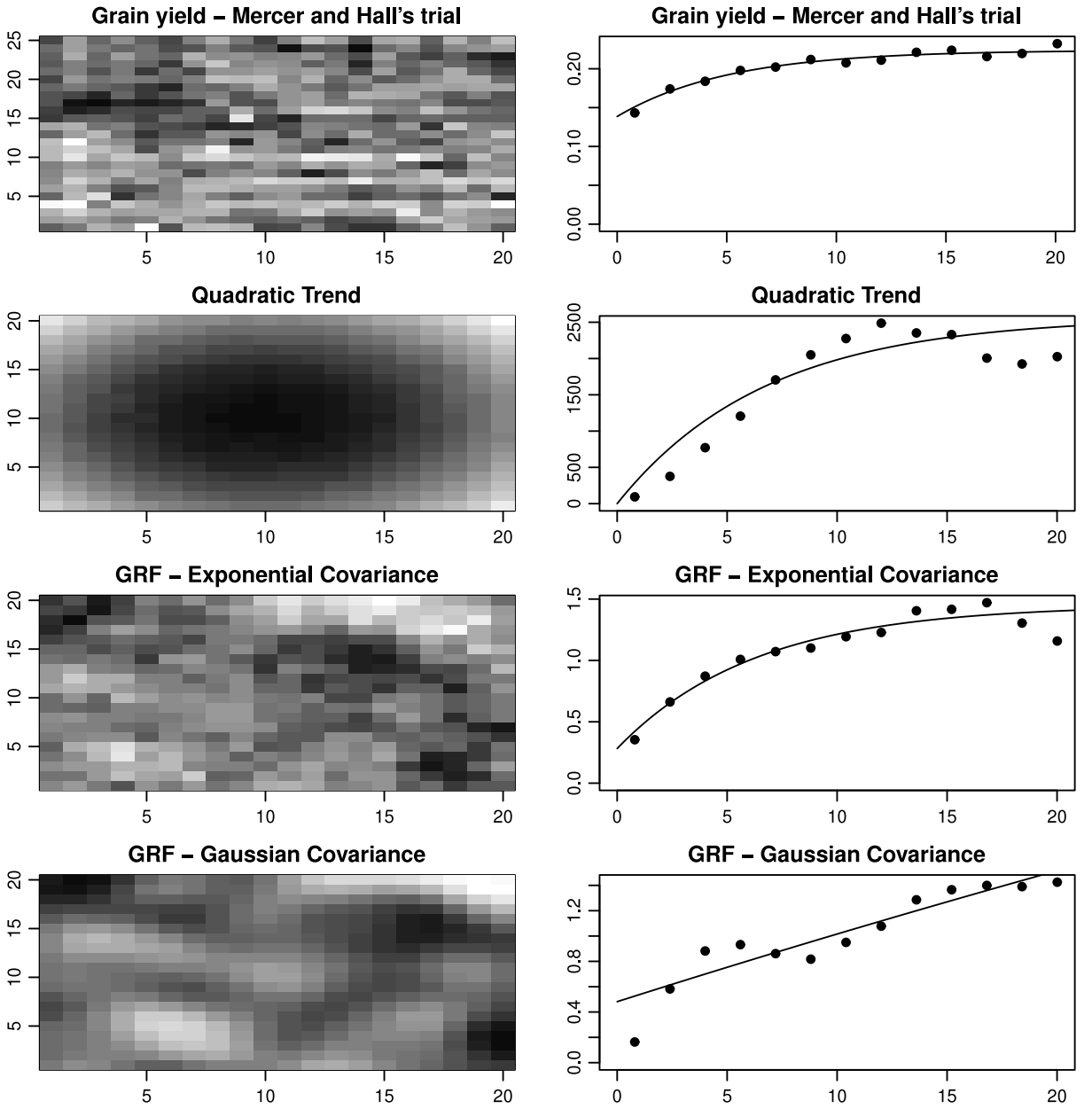

Fig. 1. Left: simulated populations (each cell in the grid is a sampling unit); Right: Empirical variogram; the classical method of moments estimator (Cressie, 1993, Chapter 2). 21 is the numerical value defining the maximum distance for the variogram. Pairs of locations separated for distance larger than this value are ignored for the variogram calculation. The correlation function is the exponential model. Value of the smoothness parameter = 0.5.

(ii) a robust to contamination of outliers estimator (Hawkins and Cressie, 1984), see also Cressie (1993, p. 175, formula 2.4.12).

$$\hat{\gamma}(d) = \frac{\frac{1}{N(d)} \sum_{(i,j)\in N(d)} \left[|Y_i - Y_j|^{1/2}\right]^4}{0.914 + 0.988/N} \tag{11}$$

(iii) a nonparametric variogram estimator as proposed by Garcia-Soidan et al. (2003). This estimator estimates a multidimensional variogram (and its first derivatives) using local polynomial kernel smoothing of linearly binned semi-variances. We have set the bandwidth parameter equal to 10 as it is done in most practical applications. For further discussion see also Fernández-Casal et al. (2003) and Fernández-Casal and Francisco-Fernández (2014).

Table 1
Maximal stratification: empirical relative bias and RMSE of the estimators. 1000 Monte Carlo replications. Mercer and Hall (MH) data; Quadratic Trend (QT) data, Gaussian random fields 1 and 2 (GRF1 and GRF2) data. Suffix denotes sample size.

	$\hat{V}_{MM}$	$\hat{V}_{RBS}$	$\hat{V}_{NP}$	$\hat{V}_{SRS}$	$\hat{V}_{SRS} \cdot c$	$\hat{V}_{SRS} \cdot I$
Relative bias						
MH	−0.012	−0.023	−0.131	0.267	0.184	−0.130
QT_{16}	0.656	0.932	3.635	2.524	2.094	0.679
QT_{25}	1.102	1.467	5.945	4.429	2.873	0.356
$GRF1_{16}$	0.171	0.103	0.775	0.878	0.740	0.738
$GRF1_{25}$	0.284	0.158	0.776	1.277	0.935	0.520
$GRF2_{16}$	−0.110	−0.163	1.573	0.611	0.429	0.422
$GRF2_{25}$	0.220	0.109	2.426	1.272	0.924	0.841
Relative RMSE						
MH	0.366	0.407	0.825	0.419	0.375	0.458
QT_{16}	0.759	1.098	4.667	2.685	2.216	0.924
QT_{25}	1.171	1.585	6.628	4.534	2.938	0.454
$GRF1_{16}$	0.523	0.516	1.834	1.059	0.929	0.993
$GRF1_{25}$	0.482	0.439	1.453	1.378	1.028	0.843
$GRF2_{16}$	0.394	0.450	2.802	0.837	0.670	0.813
$GRF2_{25}$	0.424	0.455	3.485	1.384	1.028	1.153

Based on the above variogram estimators we propose then three different variance estimators which we will denote respectively as $\hat{V}_{MM}$, $\hat{V}_{RBS}$ and $\hat{V}_{NP}$.

3.2. Results

Tables 1 and 2 report the simulation results respectively for the case of maximal stratification and two-dimensional systematic sampling. In both tables the relative bias and relative square root of the MSE are reported, i.e. for the relative bias

$$\frac{\hat{E}[\hat{V}(\hat{Y}_d)] - V(\hat{Y}_{HT})}{V(\hat{Y}_{HT})} \tag{12}$$

and for the relative RMSE,

$$\frac{\sqrt{\hat{E}[\hat{V}(\hat{Y}_d) - V(\hat{Y}_{HT})]^2}}{V(\hat{Y}_{HT})} \tag{13}$$

where, following the notation and results introduced in Sections 2 and 3, $V(\hat{Y}_{HT})$ is the true variance of the HT estimator which can be calculated because the inclusion probabilities for each sample design are known; it becomes then the benchmark for our procedures. $\hat{E}[\hat{V}(\hat{Y}_d)]$ indicates the mean obtained in the simulation runs by the different estimation strategies under each sample design; the operator $\hat{E}$ should be read as E in the case of systematic sampling as all possible samples have been considered.

As far as maximal stratification is concerned, from the first part of Table 1 (relative bias) the only circumstances of underestimation of $V(\hat{Y}_{HT})$ are in the case of MH data (weak spatial auto-correlation) and when, for small sample size, there is a wrong specification of the auto-correlation function for variogram estimator ($GRF2_{16}$). In both cases the size of the relative bias is quite small compared to that of other estimators.

In all other cases our estimators $\hat{V}_{MM}$ and $\hat{V}_{RBS}$ have a positive relative bias always smaller than that of other estimators. An exception is $\hat{V}_{NP}$; the choice of the window parameter may get a too high flexibility in the variogram estimator with consequent high variability in variance estimates. We will not pursue fine tuning of the window parameter in this context and this case will not be considered any more in our analysis.

Next, note from Table 1 that the relative RMSE of the proposed estimators is always smaller of the estimators of D'Orazio (2003) but the case of QT_{25}. Probably the Moran's index correction better captures the high positive correlation present in the population.

In order to facilitate interpreting the figures in Tables 1 and 2, the values of the relative RMSE of ours and the estimators suggested by D'Orazio have also been compared with the relative RMSE of the SRS design: Tables 3 and 4, for an estimator $\hat{V}$, report the values

$$100 \times \left(1 - \frac{relRMSE_{\hat{V}}}{relRMSE_{SRS}}\right) \tag{14}$$

to measure the efficiency gains of spatial estimators with respect to SRS (Dickson et al., 2014).

Table 2
Systematic sampling: population relative bias and RMSE of the estimators. Mercer and Hall (MH) data; Quadratic Trend (QT) data, Gaussian random fields 1 and 2 (GRF1 and GRF2) data. Suffix denotes sample size.

	$\hat{V}_{MM}$	$\hat{V}_{RBS}$	$\hat{V}_{NP}$	$\hat{V}_{SRS}$	$\hat{V}_{SRS}\cdot c$	$\hat{V}_{SRS}\cdot I$
Relative bias						
MH	0.019	0.063	−0.350	0.279	0.160	−0.237
QT_{16}	12.110	15.527	25.566	27.778	20.541	6.978
QT_{25}	1.255	2.352	0.719	4.662	2.471	0.071
$GRF1_{16}$	3.278	3.431	2.915	5.361	4.939	4.994
$GRF1_{25}$	−0.063	0.016	−1.255	0.506	0.282	0.058
$GRF2_{16}$	15.029	15.182	20.099	19.851	17.609	17.698
$GRF2_{25}$	2.808	2.170	0.442	3.870	3.020	2.496
Relative RMSE						
MH	0.411	0.528	1.641	0.462	0.389	0.506
QT_{16}	12.340	15.830	30.788	28.487	20.702	6.994
QT_{25}	1.302	2.401	2.823	4.739	2.482	0.088
$GRF1_{16}$	3.754	4.078	6.416	5.799	5.411	5.593
$GRF1_{25}$	0.268	0.327	1.820	0.646	0.425	0.427
$GRF2_{16}$	16.717	16.547	41.856	22.088	19.483	20.499
$GRF2_{25}$	2.894	2.380	7.279	4.117	3.135	2.950

Table 3
Efficiency gain for maximal stratification design. 1000 Monte Carlo replications. Mercer and Hall (MH) data; Quadratic Trend (QT) data, Gaussian random fields 1 and 2 (GRF1 and GRF2) data. Suffix denotes sample size.

Efficiency gain				
	$\hat{V}_{MM}$	$\hat{V}_{RBS}$	$\hat{V}_{SRS}\cdot c$	$\hat{V}_{SRS}\cdot I$
MH	12.66	2.86	15.80	−9.41
QT_{16}	71.75	59.13	27.33	75.45
QT_{25}	74.18	65.04	47.64	98.14
$GRF1_{16}$	50.67	51.30	6.70	3.56
$GRF1_{25}$	65.00	68.15	34.25	33.87
$GRF2_{16}$	52.88	46.21	11.79	7.20
$GRF2_{25}$	69.39	67.16	23.85	28.34

Table 4
Efficiency gain for two-dimensional systematic design. Mercer and Hall (MH) data; Quadratic Trend (QT) data, Gaussian random fields 1 and 2 (GRF1 and GRF2) data. Suffix denotes sample size.

Efficiency gain				
	$\hat{V}_{MM}$	$\hat{V}_{RBS}$	$\hat{V}_{SRS}\cdot c$	$\hat{V}_{SRS}\cdot I$
MH	11.01	−14.19	15.80	−9.41
QT_{16}	56.68	44.43	27.33	75.45
QT_{25}	72.52	49.33	47.64	98.14
$GRF1_{16}$	35.27	29.68	6.70	3.56
$GRF1_{25}$	58.53	49.35	34.25	33.87
$GRF2_{16}$	24.32	25.09	11.79	7.20
$GRF2_{25}$	29.71	42.18	23.85	28.34

Note from Table 3 that for maximal stratification, notwithstanding the good performance of $\hat{V}_{SRS}\cdot I$, the proposed estimators bring to considerable efficiency gains with respect to SRS: an average efficiency gain of 65% for $\hat{V}_{MM}$ against an average efficiency gain of 40% for $\hat{V}_{SRS}\cdot I$.

Examining the case of spatial systematic sampling in Table 2 note that the high values of relative bias and relative RMSE concern the small sample case ($n = 16$). The problem reduces substantially in the case $n = 25$. Inspection of Table 2 confirms the good performance of $\hat{V}_{SRS}\cdot I$ in the case of heavily concentrated populations (QT). In all other cases the estimators proposed here perform better. The difference in efficiency gain between $\hat{V}_{SRS}\cdot I$ and $\hat{V}_{MM}\cdot I$ is now smaller (respectively 40% and 45%) but remains in favor of the latter.

4. Conclusions

This paper suggests using a parametric or non-parametric variogram estimator in a model-based variance estimation in spatial surveys. A natural justification for this approach, as discussed in Section 2.2, stems from the analogies of the variogram and the expected variance of the HT estimator in the case of equal selection probabilities of the first

and second order. From this, extensions to other cases of interest in practical applications, specifically two-dimensional systematic sampling and maximal stratification, are derived. For these two survey strategies, simulation results show that the variogram-based estimators outperform alternative estimators in several cases and indeed they also have a good performance when the spatial correlation between units is low. Theoretical results show the consistency of the suggested estimators.

Appendix

With the notation $X_n = O_p(a_n)$ it is meant that, for any $\varepsilon > 0$ there exists a finite M such that $P(|X_n/a_n| > M) < \varepsilon\ \forall n$ and $X_n = o_p(a_n)$ meaning that, for any $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n/a_n| > \varepsilon) = 0$.

Proof of Proposition 1. We have

$$\mathrm{E}[\mathrm{Var}(\hat{Y}_{HT})] = \sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij})E\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2; \tag{15}$$

note that we can take the following estimate

$$\begin{aligned} E\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 &= E\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} - \frac{Y_j}{\pi_i} + \frac{Y_j}{\pi_i}\right)^2 \\ &= E\left[\frac{1}{\pi_i^2}(Y_i - Y_j)^2 + \left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)^2 Y_j^2 - 2\frac{1}{\pi_i}\left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)(Y_i - Y_j)Y_j\right]^2. \end{aligned} \tag{16}$$

Indicating with $\mathrm{Cov}(Y_i, Y_j) = C(i-j)$ (the covariogram) and exploiting the relationships $E(Y_i^2) = C(0) + \mu^2$, $E(Y_iY_j) - E(Y_j^2) = C(i-j) - C(0)$, $2\gamma(i-j) = 2(C(0) - C(i-j))$, the above equation becomes

$$\begin{aligned} E\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 &= \frac{1}{\pi_i^2}2\gamma(i-j) + \left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)^2 (C(0) + \mu^2) \\ &\quad + \frac{1}{\pi_i}\left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)2(C(0) - C(i-j)) \\ &= 2\gamma(i-j)\left(\frac{1}{\pi_i^2} + \frac{1}{\pi_i}\left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)\right) + \left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)^2 (C(0) + \mu^2) \\ &= 2\frac{\gamma(i-j)}{\pi_i\pi_j} + \left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right)^2 (C(0) + \mu^2). \end{aligned} \tag{17}$$

Noting that $C(0) + \mu^2 = \sigma^2$, substituting the result in (15) and exploiting the definition of $g_1(\pi, d)$ and $g_2(\pi)$ we finally obtain (3). □

Proof of Proposition 2. If $\hat{\gamma}(d)$ is consistent for $\gamma(d)$ then (9) is consistent for the estimated variance $\mathrm{E}(\mathrm{Var}(\bar{Y}_{st})) = \sum_{h=1}^{H}(1 - N_h^{-1})W_h^2\mathrm{E}(S_h^2)$. Lemmas 1 and 2, providing convergence rates, show that $\mathrm{Var}(\mathrm{Var}(\bar{Y}_{st})) = o_p(1)$ as $N \to \infty$ and as long as $H = o(N)$. It follows that $|\mathrm{E}(\mathrm{Var}(\bar{Y}_{st})) - \mathrm{Var}(\bar{Y}_{st})| = o_p(1)$ for $N \to \infty$ and the result of the proposition follows. □

Lemma 1. *Let $\{Y_t, t \in T_N\}$ satisfying Assumptions 1 and 2. Let T_{N_1} and T_{N_2} be two non-overlapping subsets of T_N. Then*

$$\mathrm{Var}(S_N^2) = O_p\left(\frac{1}{N}\right) \qquad \mathrm{Cov}(S_{N_1}^2, S_{N_2}^2) = O_p\left(\frac{1}{\sqrt{N_1}\sqrt{N_2}}\right). \tag{18}$$

Proof. Defining $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ to be respectively the (constant) mean vector and the covariance matrix of $\{Y_t, t \in T_N\}$, we can obtain an upper bound for $\mathrm{Var}(S_N^2)$ from Theorem 2 in Knautz and Trenkler (1995) as

$$\mathrm{Var}(S_N^2) \le (\mu_4 - 1)(1 - N)^{-2}\sum_{i=1}^{N-1}\lambda_i^2 \quad (\mu_4 \ge 1) \tag{19}$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$ are the eigenvalues of $\boldsymbol{\Sigma}_N$. Then $\mathrm{Var}(S_N^2) = O_p\left(\frac{1}{N}\right)$ if $\mu_4 < \infty$, true by assumption, and if

$$(1 - N)^{-1}\sum_{i=1}^{N-1}\lambda_i^2 < \infty. \tag{20}$$

To devise an easy way to check whether condition (20) holds, let us resort to matrix norm theory which tells us that, if $\|\boldsymbol{\Sigma}_N\|$ denotes a subordinate norm of $\boldsymbol{\Sigma}_N$, then $|\lambda_1| \leq \|\boldsymbol{\Sigma}_N\|$ where one can take $\|\boldsymbol{\Sigma}_N\|_\infty$, i.e. the max row sum of $\boldsymbol{\Sigma}_N$. Note next that the max row sum of $\boldsymbol{\Sigma}_N$ is $\sum_{i,j} C(i-j)$ which is finite if $\{Y_t, t \in T_N\}$ satisfies Assumption 2.

The fact that $\mathrm{Cov}(S^2_{N_1}, S^2_{N_2}) = O_p\left(\frac{1}{\sqrt{N_1}\sqrt{N_2}}\right)$ follows by an application of the Cauchy–Schwarz inequality. □

Lemma 2. *Let $N_h = N/H$, then, as $N \to \infty$ and under Assumptions 1 and, $\mathrm{Var}(\mathrm{Var}(\bar{Y}_{st})) = O_p\left(\frac{1}{N}\right)$ if H is fixed, $\mathrm{Var}(\mathrm{Var}(\bar{Y}_{st})) = o_p\left(\frac{1}{N^2}\right)$, if $H = o(N)$.*

Proof. Exploiting the results of Lemma 1,

$$\begin{aligned}
\mathrm{Var}(\mathrm{Var}(\bar{Y}_{st})) &= \sum_{h=1}^{H}\sum_{k=1}^{H}(1-N_h^{-1})(1-N_k^{-1})\left(\frac{N_h}{N}\right)^2\left(\frac{N_k}{N}\right)^2 \mathrm{Cov}(S_h^2, S_k^2) \\
&= \sum_{h=1}^{H}\sum_{k=1}^{H}(1-N_h^{-1})(1-N_k^{-1})\left(\frac{N_h}{N}\right)^2\left(\frac{N_k}{N}\right)^2 O_p\left(\frac{1}{\sqrt{N_h N_k}}\right) \\
&= \frac{1}{H^2}\left(1-\frac{N}{H}\right)^2 O_p\left(\frac{H}{N}\right) \qquad (21)
\end{aligned}$$

where in the last line we have used the simplifying assumption that $N_h = N_k = N/H$. One can see that the dominating term in the above expression, as $N \to \infty$ is of order $O_p(1/N)$ if H is fixed, while it is of order $o_p(1/N^2)$ if the number of strata H is allowed to grow with the population dimension N at rate $H = o(N)$. □

References

Bartolucci, F., Montanari, G.E., 2006. A new class of unbiased estimators of the variance of the systematic sample mean. J. Statist. Plann. Inference 136 (4), 1512–1525.
Benedetti, R., Piersimoni, F., Bee, M., Espa, G. (Eds.), 2010. Agricultural Survey Methods. Wiley, New York.
Benedetti, R., Piersimoni, F., Postiglione, P., 2015. Sampling Spatial Units for Agricultural Surveys. Springer, Berlin.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 192–236.
Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Wiley, New York.
Cochran, W.G., 1977. Sampling Techniques. Pion, London.
Cressie, N.A.C., 1993. Statistics for Spatial Data. John Wiley & Sons, New Jersey.
Dickson, M.M., Benedetti, R., Giuliani, D., Espa, G., 2014. The use of spatial sampling designs in business surveys. Open J. Stat. 4, 345–354.
Diggle, P.J., Ribeiro Jr., P.J., 2007. Model–Based Geostatistics. Springer, New York.
D'Orazio, M., 2003. Estimating the variance of the sample mean in two-dimensional systematic sampling. J. Agric. Biol. Environ. Stat. 8, 280–295.
Fernández-Casal, R., Francisco-Fernández, M., 2014. Nonparametric bias-corrected variogram estimation under non-constant trend. Stoch. Environ. Res. Risk Assess. 28, 1247–1259.
Fernández Casal, R., González Manteiga, W., Febrero Bande, M., 2003. Space–time dependency modeling using general classes of flexible stationary variogram models. J. Geophys. Res. 108, 8779. http://dx.doi.org/10.1029/2002JD002909.
Fewster, R.M., 2011. Variance estimation for systematic designs in spatial surveys. Biometrics 67, 1518–1531.
Fuller, W.A., 2009. Sampling Statistics. John Wiley & Sons, New Jersey.
Garcia-Soidan, P.H., González-Manteiga, W., Febrero-Bande, M., 2003. Local linear regression estimation of the variogram. Statist. Probab. Lett. 64, 169–179.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press on Demand.
Grafström, A., Tillé, Y., 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24, 120–131.
Gregoire, T.G., Valentine, H.T., 2007. Sampling Strategies for Natural Resources and the Environment. Chapman & Hall, New York.
Hawkins, D.M., Cressie, N.A.C., 1984. Robust kriging – a proposal. J. Int. Ass. Math. Geol. 16, 3–18.
Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London.
Knautz, H., Trenkler, G., 1995. Bounds for bias and variance of S^2 under dependence. Scand. J. Statist. 22, 121–128.
Leonenko, N., Taufer, E., 2013. Disaggregation of spatial autoregressive processes. Spat. Stat. 3, 1–20.
Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71, 135–146.
Matheron, G., 1971. The Theory of Regionalised Variables and its Applications. In: École National Supérieure des Mines, vol. 5.
Mercer, W.B., Hall, A.D., 1911. The experimental error of field trials. J. Agric. Sci. 4, 107–132.
Opsomer, J.D., FranciscoFernández, M., Xiaoxi, Li., 2012. Modelbased nonparametric variance estimation for systematic sampling. Scand. J. Statist. 39, 528–542.
Patankar, V.N., 1954. The goodness of fit of the frequency distribution obtained from stochastic processes. Biometrika 41, 450–462.
Ripley, B.D., 1981. Spatial Statistics. John Wiley & Sons, New York.
Särndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer, New York.
Stevens Jr., D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. J. Amer. Statist. Assoc. 99, 262–278.
Tan, Kim H., 2005. Soil Sampling, Preparation, and Analysis. CRC Press.
Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46, 234–240.
Wang, J.F., Christakos, G., Hu, M.G., 2009. Modeling spatial means of surfaces with stratified nonhomogeneity. IEEE Trans. Geosci. Remote Sens. 47 (12), 4167–4174.
Wang, J., Haining, R., Cao, Z., 2010. Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int. J. Geogr. Inf. Sci. 24 (4), 523–543.
Wang, J.F., Haining, R., Liu, T.J., Li, L.F., Jiang, C.S., 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. Environ. Plann. A 45 (10), 2515–2534.
Wang, J.F., Stein, A., Gao, B.B., Ge, Y., 2012. A review of spatial sampling. Spat. Stat. 2, 1–14.
Wang, J.F., Zhang, T.L., Fu, B.J., 2016. A measure of spatial stratified heterogeneity. Ecol. Indic. 67, 250–256.
Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–449.
Wolter, K.M., 2007. Introduction to Variance Estimation. Springer-Verlag Inc., New York.