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: