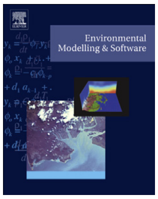
Environmental Modelling & Software 61 (2014) 87–97

Contents lists available at ScienceDirect

Environmental Modelling & Software

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



Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data

Linda Aune-Lundberg, Geir-Harald Strand*

Norwegian Forest and Landscape Institute, P.O.Box 115, N-1431 Ås, Norway

A R T I C L E   I N F O

Article history:
Received 3 April 2014
Received in revised form
26 June 2014
Accepted 3 July 2014
Available online

Keywords:
Systematic sample
Spatial sampling
Uncertainty
Variance Land cover
Land use
Area frame survey
Spatial autocorrelation

A B S T R A C T

Systematic sampling is more precise than simple random sampling when spatial autocorrelation is present and the sampling effort is equal, but there is no unbiased method to estimate the variance from a systematic sample. The objective of this paper is to assess selected variance estimation methods and evaluate the influence of spatial structure on the results. These methods are treated as models and a complete enumeration of Norway was used as the modeling environment. The paper demonstrates that the advantage of systematic sampling is closely related to autocorrelation in the material, but also that the improvement is influenced by periodicity and drift in the variables. Variance estimation by stratification with the smallest possible strata gave the best overall results but may underestimate the variance when spatial autocorrelation is absent. Treating the sample as a simple random sample is a safe and conservative alternative when spatial autocorrelation is absent or unknown.

1. Introduction

Spatial sampling surveys fill an important gap between the traditional, labor-intensive wall-to-wall field survey and the efficient, but in many cases rather inaccurate mapping by remote sensing (Wyatt, 2000; Verburg et al., 2011). The approach is used from the global down to the sub-national level. The Food and Agriculture Organization of the United Nations used systematic sampling together with satellite remote sensing for their Global Forest Resources Assessment 2010 (FAO, 2010). This approach reduced the amount of image processing and allowed FAO to involve national experts who revised the sample areas. The combination of field inventories and systematic sampling was also chosen when the European statistical agency (Eurostat) developed the LUCAS (Land use/cover area frame survey) program, carried out in the EU countries (Eurostat, 2003; Martino and Fritz, 2008). The Norwegian (Dramstad et al., 2002) and Swedish (Ståhl et al., 2011) landscape monitoring programmes both rely on area frame surveys where aerial photo interpretation is supplemented with observations from field inventories. Norway has also implemented a national area frame survey of land cover and outfield land resources (Strand, 2013). Spatial sampling methods are furthermore used in the Norwegian (Tomter et al., 2010), Swedish (Axelsson et al., 2010) and Finnish (Tomppo and Tuomainen, 2010) National Forest Inventories. The sampling approach allows these surveys to employ field observations and interpretation of high resolution imagery for large areas within acceptable budgets.

Spatial sampling surveys can be implemented following a number of different sampling strategies (Wang et al., 2012). Two of the most common are simple random sampling and systematic random sampling. Systematic random sampling is known from statistical theory to produce more precise estimates, in the spatial context and under certain conditions, than simple random sampling because the sampling units are distributed more evenly across the sampled area (Bellhouse and Sutradhar, 1988; Dunn and Harrison, 1993; D'Orazio, 2003; Ambrosio et al., 2004). This is an advantage when nearby sampling units show a high degree of positive correlation (Cochran, 1977; Flores et al., 2003), as often is the case with land use/land cover data (Legendre, 1993).

Systematic samples do have their limitations in situations with systematic variation in the landscape itself, appearing e.g. as wave or chessboard like structures (Fattorini et al., 2006). Systematic sampling also makes it more difficult to adapt to budget changes during a survey (Stehman, 2009). The overall notion is, however, that systematic sampling more often than not is found to be an efficient sampling strategy for land cover and other land resource surveys (Thompson, 2002; Stehman, 2009).

---

* Corresponding author. Tel.: +47 64 94 96 99; fax: +47 64 94 80 01.
E-mail address: ghs@skogoglandskap.no (G.-H. Strand).

http://dx.doi.org/10.1016/j.envsoft.2014.07.001

The advantage of systematic sampling does, however, come with a hitch. This sampling method can produce more precise estimates than simple random sampling, but there is no unbiased estimation method for calculation of the uncertainty and documentation of the higher precision in these surveys. The reason is that the systematic sampling design is using a single random starting point where only one unit is drawn randomly. The other units are spaced from each other at a fixed distance (Madow and Madow, 1944). This design can be described as drawing a single "cluster" of regularly spaced individuals. The sampling unit is the cluster and the sample size is $n = 1$ (Thompson, 2002). As a consequence, it is not possible to use ordinary variance estimation methods since they require a denominator of $n - 1$.

There have been attempts to provide unbiased estimation of variance in systematic samples by combining repeated systematic samples with several starting points (Koop, 1971). The approach suggested by Koop with a few replicates (for example two or three starting points chosen at random) is unbiased but unstable (the variance of the estimated variance is large). Other attempts use stratification (Gautschi, 1957) or a mixture of systematic and simple random sampling (Zinger, 1980; Wu, 1984). All these methods rely on drawing more than one single systematic sample, which is fine in an experimental situation but rarely possible in applied large-scale surveys in forestry, land use/land cover studies or ecology.

The normal approach for handling a systematic sample is to disregard the fact that the systematic sample is a cluster sample and compute the variance using the estimators intended for simple random sampling (Milne, 1959; Cochran, 1977; Wolter, 1984, 2007). This approach results in a biased and in many cases significantly overestimated result (Matèrn, 1960; Dunn and Harrison, 1993; Särndal et al., 2003), and the benefit from lower variance in systematic samples is therefore hidden (Fewster et al., 2009).

Alternative approaches using traditional variance estimation combined with a local indicator are demonstrated by e.g. Matèrn (1947), Wolter (2007) and Gallego and Delincé (2010). The principle of the local variance estimation methods is to treat neighboring observations as a pseudo-stratum. The strata can be overlapping or non-overlapping. The variation within these strata replaces the usual deviation from the overall mean in the traditional simple random sampling variance estimation method, resulting in a least biased estimate of the variance (Matèrn, 1960; Wolter, 2007). The advantage of the local variance estimation method is that it takes the spatial ordering into account and thus also the autocorrelation.

A local variance estimation method is currently used for estimation of the variance of the mean in the Finish National Forest inventory (Tomppo and Heikkinen, 1999). Likewise, Gallego and Delincé (2010) used a local estimator based on the eight nearest neighbors to each sampling point for variance estimation of the LUCAS surveys. These methods reportedly demonstrate promising results for variance estimation in applied systematic random sampling surveys. Tests involving completely enumerated population have been carried out in ecology (Aubry and Debouzie, 2000) but were limited to simple processes and small areas. Rigorous testing on real land use/land cover data is rarely reported. Only a few studies (Dunn and Harrison, 1993; D'Orazio, 2003; Opsomer et al., 2012) use real land use/land cover or forestry data and a complete enumeration of a landscape (although of restricted size) for validation. There is also a lack of examples showing how different variance estimation methods behave in situations with different spatial structure and over a range of different land use and land cover types. Finally, the literature is remarkably vague with respect to precisely how the proposed methods are implemented. The programmer is therefore left with a number of open questions when trying to implement the methods discussed in the literature in an operative environment.

The challenge described here can be approached as a need for model evaluation. At the basic level, a statistical sample – with its sampling units and selected features – is a model of an environment. The assessment of how well the sample reflects the population is a question of model performance and the choice between a simple random sample and a systematic sample is, in this context, a choice between two different models. Furthermore, a situation arises when systematic sampling has been chosen where the uncertainty of the resulting statistical estimators has no (known) mathematical solution. It is therefore necessary to develop and apply indicators to describe the uncertainty. These indicators are also models and the evaluation of alternative indicators is a study and assessment of model performance.

The purpose of this study is clearly not to break new ground in the field of spatial statistics. The relevant theory is well established. Our purpose is instead to examine estimation methods for variance calculation on different land use/land cover types in a survey by applying methods proposed for the more general characterization of the performance of environmental models (Bennett et al., 2013). The justification is partly a need for an empirical demonstration in order to explain the advantage of systematic sampling to the wider land monitoring community, partly to arrive at an applicable method for local variance estimation, which can be implemented in the setting of an operational land monitoring program. We use a complete enumeration of an extended (in our case national) dataset, which acts as a pseudo-truth. This dataset includes a combination of land use/land cover types with heterogeneous spatial structure covering a credible range of real-world situations.

The research questions examined in this study are: (1) Is the simple random sampling variance estimation method always a conservative estimate of the variance for two-dimensional systematic random samples?; (2) Does local variance estimation methods form a more precise estimate of the variance than the simple random sampling method?; (3) How do the different local estimation methods compare?; and (4) How are the results influenced by the spatial structure and distribution of the different land use/land cover types?

2. Material and methods

2.1. Material

The material used in the study consist of a digital land use/land cover map of Norway (AR50; cartographic scale 1:50,000) with seven land use/land cover classes listed in Table 1. The spatial units of AR50 are polygons and the minimum mapping unit is 1.5 ha with a geometric accuracy of 20 m. AR50 is available on Internet for viewing and downloading (http://kilden.skogoglandskap.no, last accessed June 25th 2014). The study area used in the analysis was the entire Norwegian mainland, totally 324,099 km^2.

The coverage of the different land use/land cover types is far from uniform, as shown in Fig. 1. Built-up and agricultural land are both marginal land use/land cover types in Norway. Built-up land covers only 0.5% of the total area and is highly dispersed. Agriculture covers 3.4% of the area but the pattern is clustered with some areas having a much higher percentage of agriculture, close to 50% around the Oslo fiord. Forest and open land are the two dominant land use/land cover types in

Table 1
Descriptive statistics (sum, population mean and population variance) for the seven land use/land cover types in the gridded version of the national land use/land cover map AR50. $N = 350{,}514$ grid cells.

Land cover class	N	Sum (km^2)	Mean μ (km^2)	Variance σ^2
1 Built-up land	350,514	1859.25	0.00530	0.002105
2 Agriculture	350,514	12,658.59	0.03611	0.013735
3 Forest	350,514	126,113.46	0.35980	0.134033
4 Open land	350,514	140,148.26	0.39984	0.171475
5 Mire	350,514	21,722.85	0.06197	0.016112
6 Snow/ice	350,514	3038.19	0.00867	0.005934
7 Water	350,514	18,559.31	0.05295	0.020069