Doubly Balanced Spatial Sampling with Spreading and Restitution of Auxiliary Totals

Anton Grafström[1] and Yves Tillé[23]
[1]Swedish University of Agricultural Sciences, Umeå, SWEDEN
[2]University of Neuchâtel, SWITZERLAND
[3]Corresponding author: Yves Tillé, e-mail: yves.tille@unine.ch

Abstract

A new spatial sampling method is proposed in order to achieve a double property of balancing. The sample is spatially balanced or well spread so as to avoid selecting neighbouring units. Moreover, the method also enables to satisfy balancing equations on auxiliary variables available on all the sampling units because the Horvitz-Thompson estimator is almost equal to the population totals for these variables. The method works with any definition of distance in a multidimensional space and supports the use of unequal inclusion probabilities. The algorithm is simple and fast.

Keywords: Balanced sampling; Pivotal method, Spatially balanced sampling; Spatial correlation.

1. Introduction

Statistical units selected from a territory are generally spatially correlated, which means that two neighboring statistical units tend to be more similar than two distant statistical units. A large set of publications are dedicated to methods of spatial sampling that takes into account spatial correlation. The most usual methods are systematic sampling, spatial stratification, Generalized Random-Tessellation Stratified (GRTS) sampling (see among others Ripley, 1981; Thompson, 1992; Stevens and Olsen, 2003, 2004; Mandallaz, 2008; Marker and Stevens, 2009).

In many survey problems, auxiliary information is available for all the units of the population of interest under the form of a census or a register. The auxiliary information can be spatial coordinates and/or any other variables related to the variable of interest. Let $U = \{1, 2, \dots, N\}$ denote the population of N units. We wish to estimate a total of some study variable y which takes a fixed value y_k for unit $k \in U$. A vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kp})^T$ of the values taken by p auxiliary variables is supposed to be known for each unit of the population. The spatial coordinates of unit k are also supposed to be known.

We aim to combine two main ideas for constructing efficient sampling designs that make the best possible use of available auxiliary information. The first main idea is the use of balanced sampling. Deville and Tillé (2004) introduced the cube method, which allows to select unequal or equal probability samples that are balanced or almost balanced on several auxiliary variables. Balanced sampling means that the Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) of the total of these auxiliary variables given by

$$\widehat{\mathbf{X}} = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k}$$

are equal or almost equal to the known totals given by

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k,$$

i.e.

$$\widehat{\mathbf{X}} \approx \mathbf{X},$$

where S denotes the random sample, and π_k the inclusion probability of unit k. Balanced sampling is very efficient when the study variable can be well approximated by a linear combination of the auxiliary variables (Nedyalkova and Tillé, 2008).

The second main idea is spatially balanced sampling, which means that the samples are well spread in the space so as to avoid selecting neighbouring units. Stevens and Olsen (2004) introduced GRTS sampling. Their method uses a specific random mapping from two-dimensional or multi-dimensional locations to one dimension. The sample is then selected by a systematic design in one dimension and mapped back to two or more dimensions. This procedure guarantees that each sample is rather well spread over the population. Lister and Scott (2009) have used space-filling curves in order to make sure that the sample locations are well spread over the space.

Grafström (2011) and Grafström et al. (2012) introduced new sampling methods that enable to select unequal probability samples that are well spread over the population. These methods are respectively called spatially correlated Poisson sampling (SCPS) and local pivotal method. Instead of a mapping, these methods use distance between units to create small joint inclusion probabilities for nearby units, forcing the samples to be well spread. An advantage of SCPS and the local pivotal method is that the use of a distance measure makes it easy to spread the sample in any number of dimensions. Spatially balanced sampling is efficient when there are spatial trends within the population (see e.g. Stevens and Olsen, 2004). Indeed, nearby locations or units usually have similarities. These similarities can be due to similar conditions in the environment. In this situation, it is efficient to make sure that the sample is well spread, i.e. it is unwise to select nearby units. Spatially balanced sampling is commonly used for natural resources, which often exhibit spatial trends.

2. Strategy for balanced sampling

If the population of interest is generated by a linear model with uncorrelated errors terms, Nedyalkova and Tillé (2008) have shown that the best model-assisted strategy is to first randomly select a balanced sample with inclusion probabilities proportional to the standard deviations of the errors and then to use the HT-estimator of total. When sampling from a territory, the units are often spatially correlated. This can be formalized by means of the following linear model:

$$y_k = \mathbf{x}_k^T \boldsymbol{\beta} + \varepsilon_k, \text{ for all } k \in U, \tag{1}$$

where $\mathbf{x}_k$ is a column vector of the values taken by the p auxiliary variables on unit k, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients. Moreover, the ε_k are random variables such that $\mathrm{E}_M(\varepsilon_k) = 0$, $\mathrm{var}_M(\varepsilon_k) = \sigma_k^2$, for all $k \in U$, and

$$\mathrm{cov}_M(\varepsilon_k, \varepsilon_\ell) = \sigma_k \sigma_\ell \rho_{k\ell}, \text{ with } k \neq \ell \in U,$$

where $\mathrm{E}_M(.)$, $\mathrm{var}_M(.)$ and $\mathrm{cov}_M(.)$ respectively denote the expectation, variance and covariance under model (1).

Usually the closer the units are, the more correlated they are. The $\rho_{k\ell}$ are thus supposed to be decreasing in function of a distance that can be computed between k and ℓ. For instance, the correlations could be written as $\rho_{k\ell} = \rho^{d(k,\ell)}$, where $d(k, \ell)$ is a distance between units k and ℓ.

Let $p(s)$ be a sampling design on the population, S be the random sample with fixed sample size n, π_k be the first order inclusion probability, $\pi_{k\ell}$ be the joint inclusion probability, and

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

be the HT-estimator of the total

$$Y = \sum_{k \in U} y_k.$$

It is possible to show that, under model (1), a very efficient sampling design in order to estimate Y by the HT-estimator consists of:

- using a balanced sampling design on the independent variable $\mathbf{x}_k$,
- avoiding the selection of neighboring units, i.e selecting a well spread sample (or spatially balanced),
- using inclusion probabilities proportional to σ_k.

Notice that the use of the HT-estimator under a design with these properties will be efficient if the population is close to a realization from the model, but maintains desirable properties like design unbiasedness and design consistency even if the model is false. The use of the HT-estimator thus guaranties the robustness against a miss-specification of the model.

3. An algorithm for spread and balanced sampling

Our aim is to propose a new algorithm to select a sample that is balanced on p auxiliary x-variables and is well spread in some space. Distance between units can be measured in other variables than the x-variables on which we balance the sample. The sampling algorithm is a mixture of the cube method (Deville and Tillé, 2004) and a generalization of the local pivotal method (Grafström et al., 2012). The basic idea is to repeatedly apply the flight phase of the cube method on a cluster of $p+1$ nearby units. When the flight phase is applied on such a cluster, the sampling outcome is decided for at least one of the units, while respecting the p balancing conditions. Since the updating of the inclusion probabilities is done locally, this procedure gives small joint inclusion probabilities for nearby units. When there are less than $p+1$ units left, for which the sampling outcome is undecided, the sample is finalized by applying the landing phase of the cube method. More precisely the procedure is described in Algorithm 1.

Algorithm 1 Algorithm for spread and balanced sampling

- $\boldsymbol{\pi}(0) = \boldsymbol{\pi}, j = 0$
- While there are at least $p+1$ units whose sampling outcome are undecided, i.e. $\#A(j) \geq p+1$, where $A(j) = \{k \in U | 0 < \pi_k(j) < 1\}$.
  1. A subset $B(j)$ of $p+1$ neighboring units is selected from $A(j)$ by means of Algorithm 2.
  2. A flight phase of the cube method is applied on the $p+1$ selected units. This flight phase transform in $B(j)$ the $\pi_k(j)$ in $\pi_k(j+1)$ and satisfies
     $$\sum_{k \in B(j)} \frac{\mathbf{x}_k}{\pi_k} \pi_k(j+1) = \sum_{k \in B(j)} \frac{\mathbf{x}_k}{\pi_k} \pi_k(j).$$
     Notice that, for this flight phase, the population of reference is $B(j)$, the balancing variables are $\pi_k(j)\mathbf{x}_k/\pi_k$ and the inclusion probabilities are $\pi_k(j)$. For the units of U that are not in $B(j)$, the values $\pi_k(j)$ remain unchanged, i.e. $\pi_k(j+1) = \pi_k(j)$.
  3. Compute $j = j+1$.
- A landing phase of the cube method is applied.

A cluster of $p+1$ nearby units can be selected by Algorithm 2. With this procedure, the sample is as well balanced as with the usual cube method. The sample is also well spread. Indeed, at each step of Algorithm 1, a decision is once and for all taken for a statistical unit. If this statistical unit is taken, the inclusion probabilities of the other units of the cluster are generally decreased because the sum remains unchanged. Likewise, if the decision consists of not taking a unit, the inclusion probabilities of the other units of the cluster are generally increased. So, the method avoids the selection of neighbors. It is difficult to give a formal proof that the samples are well spread in the general case. However, with only an intercept used as auxiliary variable the new method coincide with the local pivotal method. For that special case Grafström et al. (2012) provided some theoretical results that supports that the resulting samples are very well spread.

Algorithm 2 Cluster selection algorithm

```
(1) Select among the undecided units (i.e. from $A(j)$), one unit $k$ randomly and then the
    $p$ closest units to unit $k$.
(2) Calculate the mean position of the $p+1$ units.
(3) Select the nearest $p+1$ units to the mean position.
(4) Repeat (2)-(3) while the sum of squares of the distances of the units of the cluster to
    their mean is decreasing.
```

Since the sampling outcome is decided for at least one unit in each step of the algorithm, there are at most $N-p$ steps until the landing phase can be applied and a sample is achieved. By using the R-package 'sampling', which includes functions for applying the flight phase and the landing phase of the cube method, this algorithm is easily implemented in R. The R code of the new method is available on demand.

4. Discussion

We have run a large set of simulations that show that the new method is more efficient than using only balanced sampling due to remaining spatial trends in residual terms. The new method is also more efficient than a design that only spreads the samples in the topographical space. Because the new method can both spread and balance the samples, it enables one to use more information than other alternatives. Hence, it performs better. The simulations also show that spreading and balanced sampling can be efficient even if the relationships between the x-variables and the study variables are not exactly linear. Even though we justify the method by using a superpopulation model, the inference is based on the sampling design. The HT-estimator will be efficient if the population is close to a realization from model (1), but the estimator maintains desirable properties like design unbiasedness and design consistency even if the model is not properly specified.

References

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Grafström, A. (2011). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference*, 142:139–147.

Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2):514–520.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Lister, A. and Scott, C. (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment*, 149:71–80.

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories.* Chapman & Hall/CRC, Boca Raton, FL.

Marker, D. A. and Stevens, Jr., D. L. (2009). Sampling and inference in environmental surveys. In *Sample surveys: design, methods and applications*, volume 29 of *Handbook of Statist.*, pages 487–512. Elsevier/North-Holland, Amsterdam.

Nedyalkova, D. and Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Ripley, B. D. (1981). *Spatial Statistics.* John Wiley & Sons.

Stevens, D. L. and Olsen, A. R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465):262–278.

Stevens, D. L. J. and Olsen, A. R. (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14(6):593–610.

Thompson, S. K. (1992). *Sampling.* Wiley, New York.