Performing Horvitz-Thompson estimation in spatial sampling: a computer-intensive perspective for complex schemes

Lorenzo Fattorini[1]

[1] Dipartimento di Metodi Quantitativi, Universita' di Siena, P.za S. Francesco 8, 53100 Siena (Italy)

Abstract: A modification of the Horvitz-Thompson estimator is proposed in complex spatial sampling. The inclusion probabilities are estimated by means of independent replications of the sampling scheme. The properties of the resulting estimator are derived.

Keywords: Independent replications; Monte Carlo methods; Sequential sampling scheme; Spatial and environmental sampling.

1 Introduction

Spatial sampling is widely applied in environmental surveys as when, for example, population units are administrative districts in a study region, or farms in an agricultural country or patches in a nature reserve. The present paper proposes the use of some sequential sampling schemes which may be suitable to handle the main problems of spatial sampling, i.e. the presence of spatial correlation as well as the presence of units of unequal sizes. Unfortunately, these proposals involve algebraic complications in the expressions of inclusion probabilities. Hence, the Horvitz-Thompson estimator is generally inapplicable. However, when dealing with with-frame populations and when the inclusion probabilities do not depend on the values of the interest variable in the population, the problem may be straightforwardly bypassed. Indeed, in these cases, the sample selection can be independently replicated for an adequate number of times, in such a way that the population values are quantified in the field only for the units in the first sample, while the remaining samples are used to estimate the first- and second-order inclusion probabilities on the basis of the proportion of times in which the units or the couples of units enter the selected samples.

2 Problems and solutions in spatial sampling

When population units are plots partitioning a study region, simple random sampling may prove to be an inefficient design. Indeed, adjacent units are

often more alike than units that are far apart, thus giving a poor contribution to the sample information. Moreover, in other situations, an opposite trend may be present in which adjacent units greatly differ from each other. These problems are well recognized in a model-based setting as positive or negative spatial autocorrelation problems. A procedure which seems to be attractive from a very practical point of view is proposed by Fattorini and Ridolfi (1997). The authors suggest modifying simple random sampling without replacement in such a way that, at each drawing, the probabilities of selecting those units that are adjacent to the previously selected ones are reduced or increased according to a prefixed factor $\beta \geq 0$. It can be proved that the design exists for any $\beta > 0$, in which case $\pi_{ij} > 0$ for all $h > j = 1, 2, ..., N$, while for $\beta = 0$, the second-order inclusion probabilities vanish for contiguous units and the design may fail to exist. Moreover, as $\beta \to \infty$, only adjacent units can be selected and second-order inclusion probabilities may vanish for very distant units.
Obviously, when units are of equal size, the variance of the population values simply measures the spatial variation of the interest variable across the study area. On the other hand, when the units are of different sizes, say $A_1, ..., A_N$, the population variance is inflated by the variation of plot size, thus rendering inadvisable the use of simple random sampling. In these cases, if the interest variable is likely to increase with size, a suitable solution may be the use of a ΠPS design in which the first-order inclusion probabilities are proportional to the unit sizes. Obviously, this may be straightforwardly performed by adopting one of the sampling schemes listed in Brewer and Hanif (1983). However, in many situations, the presence of unequal-sized units may concur with the presence of spatial autocorrelation. Thus, in order to take into account both these problems, Barabesi *et al.* (1997) suggest modifying the sampling scheme proposed by Skalski (1994) in which, at each drawing the probability of selecting a unit is proportional to its size (the so-called PPS schemes). Once again the modification is performed by reducing or increasing the selection probabilities of those units that are adjacent to the previously-selected ones according to a factor $\beta \geq 0$.
Fattorini and Ridolfi (1997) give some guidelines for the β choice, showing that, in presence of a marked positive or negative autocorrelation, the efficient choices are $\beta = 0$ or $\beta = \infty$, respectively.

3 The estimation of inclusion probabilities

Being based on n sequential drawings, both the Fattorini and Ridolfi (1997) and Barabesi *et al.* (1997) schemes are easy to implement in practice. However the Horvitz-Thompson estimator is inapplicable even for moderate values of N and n, since the computation of first- and second-order inclusion probabilities involves enumerating all the possible samples and all the orderings in which the units enter the sample.

When an explicit derivation of the first-order inclusion probabilities π_1, π_2, ..., π_N is prohibitive, the Horvitz-Thompson estimator, say $\hat{T}$, cannot be obtained. However, if the inclusion probabilities do not depend on the population values $y_1, y_2, ..., y_N$, these probabilities may be suitably estimated using a Monte Carlo method. Indeed, in this case, as proposed by Fattorini (2006), $M + 1$ samples of size n can be independently selected from the population frame. Subsequently, the values of the survey variable are quantified in the field only for the units in the first sample S, while the remaining M samples, say S_1,S_2, ...,S_M, are just used to estimate the inclusion probabilities. Accordingly, a suitable estimator of π_j which turns out to be invariably positive is given by

$$p_j = \frac{1 + X_j}{M + 1}, j = 1, 2, ..., N$$

where X_j is the number of times unit j enters the M samples. Since p_j constitutes a consistent ($M \to \infty$) estimator of π_j, a very natural modification of $\hat{T}$ may be

$$\hat{T}_M = \sum_{j \in \mathsf{S}} \frac{y_j}{p_j}$$

As to the statistical properties of $\hat{T}_M$, it is at once apparent that it converges almost surely to $\hat{T}$ as M increases. Moreover, Fattorini (2006) proves that turns out to be asymptotically ($M \to \infty$) equivalente to $\hat{T}$, since it is asymptotically unbiased with a mean squared error which converges to the variance of $\hat{T}$.

As to the estimation of the variance of $\hat{T}_M$, a suitable estimator of π_{ij} is given by

$$p_{ij} = \frac{X_{ij} + 1}{M + 1}, h > j = 1, 2, ..., N$$

where X_{ij} now denotes the number of times units j and h enter the M samples jointly. Then, providing that $\pi_{ij} > 0$ for each $h > j = 1, 2, ..., N$, the quantity

$$\nu_M^2 = \sum_{j \in \mathsf{S}} y_j^2 (\frac{1}{p_j^2} - \frac{1}{p_j}) + 2 \sum_{j \in \mathsf{S}} \sum_{h > j} y_j y_h (\frac{1}{p_j p_h} - \frac{1}{p_{jh}})$$

constitutes an asymptotically unbiased estimator of var($\hat{T}_m$). On the other hand, if $\pi_{jh} = 0$ for some $h > j = 1, 2, ..., N$, the quantity s_M^2/n, where $t_j = n y_j / p_j$ and

$$s_M^2 = \frac{1}{n - 1} \sum_{j \in \mathsf{S}} (t_j - \hat{T}_M)^2$$

tends to be an asymptotically conservative estimator of var($\hat{T}_m$).

As to the guidelines for choosing a suitable value of M in order to have a negligible loss of accuracy and efficiency with respect to the performance

of $\hat{T}$, Fattorini (2006) proves by simulation studies that the efficiency and accuracy losses arising from the use of empirical inclusion probabilities are usually negligible for $M = 10^6$.

References

Barabesi, L., Fattorini, L. and Ridolfi, G. (1997). Two-phase surveys of elusive populations. In: *Proceedings of the Statistics Canada Symposium 97: New Direction in Surveys and Censuses*, pp 285-287. Ottawa: Statistics Canada.

Brewer, K.R.W., Hanif, M. (1983). *Sampling with unequal probabilities*. New York: Springer-Verlag.

Fattorini, L., Ridolfi, G. (1997). A sampling design for areal units based on spatial variability. In: *Metron*, **55**, 59-72.

Fattornini, L. (2006). Applying the Horvitz-Thompson criterion in complex design: a computer intensive perspective for estimating inclusion probabilities. *Biometrika*, **93**, forthcoming.

Skalsky, J.R. (1994). Estimating wildlife populations based on incomplete area surveys. *Wildl. Soc. Bull.*, **22**, 192-203.