On the use of auxiliary information in spatial sampling

Chiara Bocci, Emilia Rocco

1. Introduction

In many fields of application it's common to be interested in spatially-related phenomena and in particular to deal with attributes which are defined on continuous spatial domains. In this framework, if the design-based approach is assumed, the attribute is usually expressed as a function $y(\mathbf{s})$ taking values on a suitable subset $\mathbf{s}$ of the plane. In the simplest case $y(\mathbf{s})$ represents the value of an attribute at the location $\mathbf{s}$. As an example, in forestal surveys $y(\mathbf{s})$ could be the amount of biomass measured in sampled sites over a forestal area; in environmental studies $y(\mathbf{s})$ could be the quantity of plastic materials collected by net tows in sampled areas over seas; etc... .

Technology development has led to a growing availability of low-cost spatial data ready-to-use, frequently derived from large scale observations (i.e. data from pervasive systems like GPS sensors, or remote sensing data from earth observation technologies). Oftentimes, these data can't directly answer specific questions posed by researchers and data users, or even if they can they are subject to measurement errors or self-selection bias. In both cases it is still necessary to rely, at least partially, on ad-hoc probabilistic surveys. On the other hand, the precision and quality of surveys estimates can be improved by using the data derived from these new sources as auxiliary information in the design phase and/or in the estimation phase.

Geographical data generally show a spatial pattern and an uneven spatial distribution over the population. In fact, usually spatial observations are not mutually independent and tend to be more similar to their neighbours. As stated by Tobler's first law of geography (Tobler, 1970): "everything is related to everything else, but near things are more related that distant things". This arises because nearby units interact with one another and tend to be influenced by the same set of natural and anthropogenic factors.

In such situations, it is well known that to estimate a mean or a total of a target variable selecting the units spatially best spread allows to collect more information and consequently provides better estimation. An important problem of sampling is thus to spread at best the sampled units in space. When, in addition to the spatial allocation, the value of one or more auxiliary variables is known for all the population units over the spatial domain, exploiting this information in the sampling design could further improve survey estimates.

A well-spread sample is usually said to be spatially balanced. Different types of spatially balanced sampling designs have been suggested in literature for sampling spatial population. Many, but not all of them, allow the use of auxiliary information, in a more or less simple way, during the units' selection procedure. For example various types of multi-phase systematic designs are used in different countries to produce National Forest Inventories for their forest monitoring programs. Tillé (2020, Chapter 8), Tillé and Wilhelm (2017), Benedetti et al. (2012) and Wang et al. (2012) give a review of the main spatial sampling methods. Since we are focusing on data that come from large scale observation (i.e. remote sensing data) to produce estimates at large scale, in the following we will focus on balanced sampling designs that can be easily implemented for big datasets.

We consider several sampling strategies, based on the spatially Balanced Sampling through Local Pivotal Method (LPM) introduced by Grafström et al. (2012), in order to identify the

Chiara Bocci, University of Florence, Italy, chiara.bocci@unifi.it, 0000-0001-8189-4445
Emilia Rocco, University of Florence, Italy, emilia.rocco@unifi.it

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

one which exploits geographical location and other sources of information to produce estimates for a spatially-related phenomenon in a more cost-efficient way. A strategy which could be globally applied by accounting for different areas characteristics in both the study and auxiliary variables, as well as for the differences in their relation. In all but one of the strategies under evaluation the sampling scheme consists of a different variation of the LPM, and therefore a single-phase non-informative sampling design is implemented. In addition, we propose an informative design which is based on a sequential use of the LPM and draws the final sample in two (or more) steps: (i) in the first step we collect an initial sample of observations on the target variable, which is used to investigate the relation between the auxiliary and study variables; (ii) then, this relation is exploited to target and tailor the subsequent sampling step; (iii) additional steps can be included by applying the procedure iteratively; (iv) finally, observations on the target variable collected in all the steps are used in the estimation process of the population mean.

The performance of the different strategies is investigated through Monte Carlo experiments by considering several scenarios, which differ in the distributions of the auxiliary and study variables and in their relation.

2. Sampling methods

Usually, in a spatial setting, the population units are plots or cells of a grid overlapping an area of interest. A value, y_i, of a variable of interest is associated with each unit $i (i = 1, ..., N)$ of the population. Moreover for each unit the spatial location $\mathbf{s}_i$, $\mathbf{s} \in \mathbf{R^2}$ is known. Here, in addition we assume to know the value x_i of an auxiliary variable for each unit of the population.

For drawing a spatial sample from such a population we decided to consider as starting point the spatially Balanced Sampling through Local Pivotal Method (LPM) introduced by Grafström et al. (2012) since it is a flexible spatially balanced design that can draw equal and unequal probability samples in multiple dimensions. Unequal probability sampling can be more efficient than equal probability sampling if there is a positive correlation between the inclusion probabilities and the response values. Additional dimensions could include any auxiliary information in addition to the spatial coordinates.

The basic idea of LPM is to avoid that units close in distance appear together in the sample. First an inclusion probability $0 < \pi_i \leq 1$ is assigned to each unit so that their sum over the population is equal to the fixed sample size. The sample is then obtained in at most N steps, where N is the population size. At each step one unit i is selected randomly from the available population and another unit j is chosen among the remaining units in the population by minimizing a distance function among them. This can be a univariate or a multivariate function that measures the distance with respect to one or more auxiliary variables, among which we can include the spatial coordinates. When all the variables are continuous the Euclidean distance is commonly used. Moreover, when multiple auxiliary variables are used, they are usually standardized or scaled in order to balance the contribution of each variable. After the selection of the unit i and j their inclusion probabilities are updated by using the following rule:

$$\begin{aligned} \text{if } \pi_i + \pi_j < 1 \text{ then } (\pi_i', \pi_j') &= \begin{cases} (0, \pi_i + \pi_j) \text{ with probability } \frac{\pi_i}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ with probability } \frac{\pi_j}{\pi_i + \pi_j} \end{cases} \\ \text{if } \pi_i + \pi_j \geq 1 \text{ then } (\pi_i', \pi_j') &= \begin{cases} (1, \pi_i + \pi_j - 1) \text{ with probability } \frac{1 - \pi_j}{2 - \pi_i - \pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ with probability } \frac{1 - \pi_i}{2 - \pi_i - \pi_j} \end{cases} \end{aligned} \tag{1}$$

As a result, in each step at least one unit is removed from the population frame, either because its probability becomes zero, and consequently it is definitely excluded from the sample, or because its probability becomes one and therefore is included in the sample. The procedure continues, updating at each step the probabilities of inclusion obtained in the previous step, until all units in the population are processed. LPM selects the units with the same probability π_is initially assigned to them, therefore the population mean can be estimated with the usual Horvitz-Thompson estimator.

The following specific LPM based sampling designs, which differ in how they exploit location and auxiliary information, have been investigated:

1. **SpatLPM**: The original formulation of the spatially balanced sampling through LPM which produces samples that are well spread in the geographic space and is based on equal inclusion probabilities.
2. **AuxLPM**: Sampling, with equal inclusion probabilities, balanced through LMP in the space spanned by the auxiliary variable.
3. **BivLPM**: Sampling, with equal inclusion probabilities, balanced through LMP in the space spanned by both the geographical coordinates and the auxiliary variable.
4. **UneqLPM**: Spatially balanced sampling through LMP with unequal inclusion probabilities π_is proportional to the auxiliary variable.
5. **StrPropAuxLPM** and **StrNeyAuxLPM**: Stratified sampling with AuxLPM design in each stratum. The area of interest is partitioned in sub-areas (strata) and then the AuxLPM is applied in each stratum. Two allocation rules are considered: Proportional and Neyman's with respect to the variance of the auxiliary variable.
6. **StrPropBivLPM** and **StrNeyBivLPM**: Stratified sampling with BivLPM design in each stratum. The same stratification designs described in the previous point, but with BivLPM applied in each stratum.
7. **SeqUneqLPM**: First an initial UneqLPM sample of size $n_0 \leq n$ (n is the final size of the sample) is selected and used to investigate the relation between the auxiliary and study variables, specifically to estimate the parameters of a generalized additive model (GAM); then the predicted values of Y are used to draw the remaining sample units with spatially balanced sampling through LMP with unequal inclusion probabilities π_is proportional to predicted values; finally an Horvitz-Thompson-type estimator is applied to produce a mean estimation that exploits the data collected in both steps.

A possible alternative to the LPM method could be the double balanced sampling of Grafström and Tillé (2012), however this sampling design is highly computationally demanding when applied to big datasets, and was unfeasible in our experiments. Conversely, LPM design has been optimized for large datasets using k-d trees (Lisic and Cruze, 2016), allowing to run our Monte Carlo experiments in a reasonable amount of time.

3. Simulation study

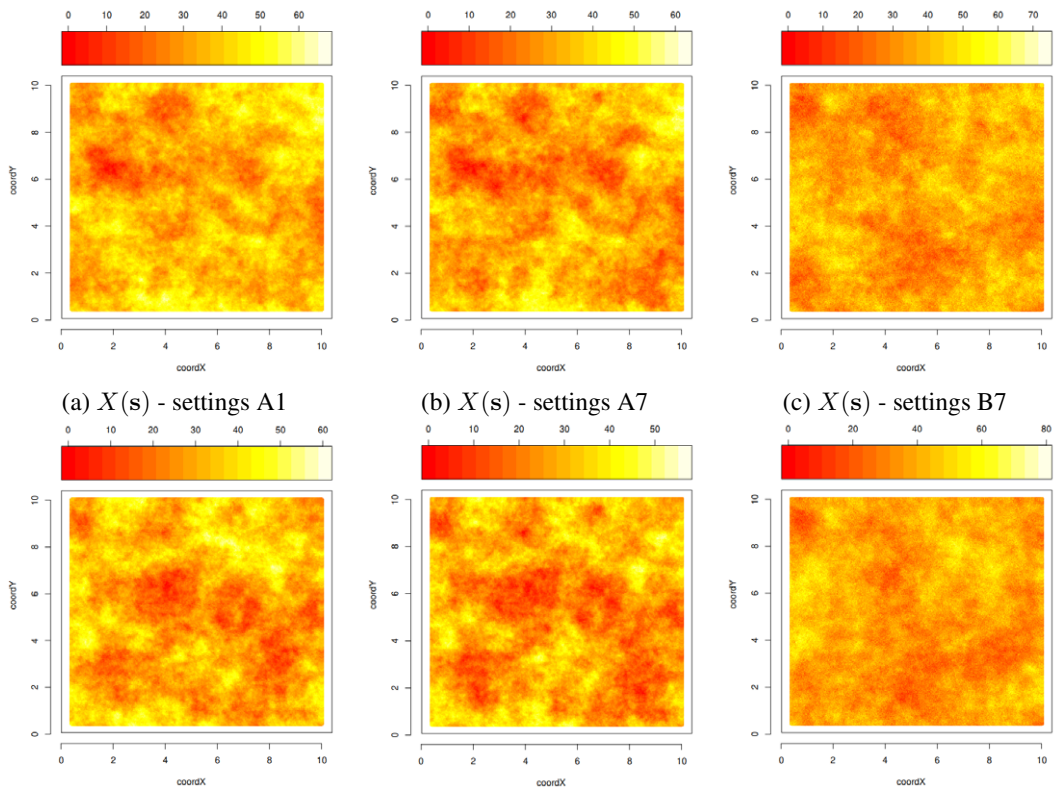
We investigate the performance of the different sampling designs through Monte Carlo experiments based on several synthetic datasets. In each of them the auxiliary (X) and response (Y) variables are drawn from a stationary bivariate spatial process $[X(\mathbf{s}), Y(\mathbf{s})]$ with $\mathbf{s} \in [0, 10]^2$ (1000×1000 grid). Following Diggle and Ribeiro (2007, Chapter 3), each bivariate spatial pro-

cess in turn is obtained as:

$$\begin{aligned} X(\mathbf{s}) &= f(a * Z_1(\mathbf{s}) + c * Z_2(\mathbf{s})) + k_1 \\ Y(\mathbf{s}) &= g(b * Z_1(\mathbf{s}) + d * Z_3(\mathbf{s})) + k_2 \end{aligned} \tag{2}$$

where:

- $Z_1(\mathbf{s}), Z_2(\mathbf{s}), Z_3(\mathbf{s}) \sim$ are independent univariate stationary Gaussian processes with an Exponential variogram with scale 1 and sill $\sigma^2 = C + C_0 = 50$, where C is the partial sill and C_0 is the nugget. For $Z_2(\mathbf{s})$ and $Z_3(\mathbf{s})$ we assume $C = 50$ and $C_0 = 0$ in all cases, while for $Z_1(\mathbf{s})$ their values vary ($C = 50, 30, 0$) to change the proportion of the co-variability that has spatial structure;
- a, b, c and d are constants whose values vary in order to obtain different correlation level and structure;
- $f(\cdot)$ and $g(\cdot)$ are transformation functions, for which we consider two choices: Identity or Exponential;
- k_1 and k_2 are adding constants to guarantee $X(\mathbf{s}) > 0$ and $Y(\mathbf{s}) > 0$.



(a) $X(\mathbf{s})$ - settings A1 (b) $X(\mathbf{s})$ - settings A7 (c) $X(\mathbf{s})$ - settings B7

(d) $Y(\mathbf{s})$ - settings A1 (e) $Y(\mathbf{s})$ - settings A7 (f) $Y(\mathbf{s})$ - settings B7

Figure 1: Variables $X(\mathbf{s})$ and $Y(\mathbf{s})$ simulated under settings A1, A7 and B7.

Overall, we present our results for 24 synthetic datasets which differ in the spatial distribution of both the study and auxiliary variables, as well as in their relation. The complete list of settings used to generate the synthetic datasets is presented in Table 1. To give a better idea of the different relations between X, Y and $\mathbf{s}$ that can be simulated in our data, Figure 1 shows

Table 1: Root mean square error, with 500 replications and sample size = 1000

Settings A: Z_1, Z_2, Z_3 with $C = 50, C_0 = 0$ $f(\cdot)$ and $g(\cdot)$ = Identity

Id	a	b	c	d	Emp.corr.	SRS	SpatLPM	AuxLPM	BivLPM	UneqLPM	SeqUneqLPM
A1	0.6	0.6	1	1	0.298	0.265	0.093	0.257	0.131	0.130	0.113
A2	0.385	1	1	0.385	0.350	0.248	0.088	0.221	0.128	0.111	0.102
A3	1	0.385	0.385	1	0.361	0.242	0.085	0.244	0.115	0.143	0.103
A4	0.82	0.82	1	1	0.424	0.295	0.103	0.250	0.149	0.138	0.119
A5	1	1	1	1	0.515	0.324	0.113	0.314	0.151	0.131	0.119
A6	1	1	0.82	0.82	0.607	0.297	0.104	0.228	0.113	0.110	0.106
A7	1	1	0.6	0.6	0.738	0.269	0.095	0.178	0.095	0.078	0.082

Settings B: Z_1 with $C = 30, C_0 = 20$ Z_2, Z_3 with $C = 50, C_0 = 0$ $f(\cdot)$ and $g(\cdot)$ = Exponential

Id	a	b	c	d	Emp.corr.	SRS	SpatLPM	AuxLPM	BivLPM	UneqLPM	SeqUneqLPM
B1	0.6	0.6	1	1	0.310	0.266	0.129	0.255	0.130	0.151	0.122
B2	0.385	1	1	0.335	0.340	0.241	0.161	0.217	0.153	0.150	0.139
B3	1	0.385	0.385	1	0.400	0.244	0.107	0.236	0.117	0.130	0.102
B4	0.82	0.82	1	1	0.437	0.295	0.156	0.264	0.147	0.140	0.136
B5	1	1	1	1	0.526	0.322	0.181	0.277	0.149	0.138	0.134
B6	1	1	0.82	0.82	0.617	0.294	0.173	0.234	0.130	0.114	0.111
B7	1	1	0.6	0.6	0.744	0.264	0.166	0.189	0.103	0.085	0.083

Settings C: Z_1 with $C = 0, C_0 = 50$ Z_2, Z_3 with $C = 50, C_0 = 0$ $f(\cdot)$ and $g(\cdot)$ = Identity

Id	a	b	c	d	Emp.corr.	SRS	SpatLPM	AuxLPM	BivLPM	UneqLPM	SeqUneqLPM
C1	0.6	0.6	1	1	0.298	0.250	0.154	0.236	0.142	0.129	0.135
C2	0.385	1	1	0.335	0.357	0.228	0.233	0.206	0.189	0.180	0.196
C2	1	0.385	0.385	1	0.345	0.233	0.114	0.240	0.115	0.159	0.102
C4	0.82	0.82	1	1	0.429	0.275	0.200	0.303	0.158	0.138	0.142
C5	1	1	1	1	0.522	0.300	0.239	0.287	0.158	0.133	0.150
C6	1	1	0.82	0.82	0.616	0.274	0.236	0.226	0.142	0.108	0.127
C7	1	1	0.6	0.6	0.747	0.247	0.234	0.167	0.105	0.078	0.099

Settings D: Z_1 with $C = 30, C_0 = 20$ Z_2, Z_3 with $C = 50, C_0 = 0$ $f(\cdot)$ and $g(\cdot)$ = Exponential

Id	a	b	c	d	Emp.corr.	SRS	SpatLPM	AuxLPM	BivLPM	UneqLPM	SeqUneqLPM
D1	0.12	0.1	0.08	0.09	0.508	0.038	0.022	0.033	0.018	0.015	0.015
D2	0.1	0.1	0.05	0.05	0.723	0.030	0.018	0.021	0.013	0.007	0.010

Settings E: Z_1, Z_2, Z_3 with $C = 50, C_0 = 0$ $f(\cdot)$ and $g(\cdot)$ = Exponential

Id	a	b	c	d	Emp.corr.	SRS	SpatLPM	AuxLPM	BivLPM	UneqLPM	SeqUneqLPM
E1	0.1	0.1	0.05	0.05	0.763	0.026	0.013	0.018	0.011	0.007	0.008

variables $X(\mathbf{s})$ and $Y(\mathbf{s})$ generated under settings A1, A7 and B7: in scenario A1 we observe a weak correlation between X and Y (equal to 0.298), with both variables strongly related with space; in both scenarios A7 and B7 the correlation between X and Y is stronger (more than 0.7), but they differ with respect to the spatial structure of the data since in scenario B7 part of the co-variability (about 40%) is not spatially related.

We choose to simulate scenarios with the different settings discussed above because when the analysis concerns a phenomenon measured at global scale it is common to observe different pattern between different areas of the globe and our aim is to find a strategy which could be globally applied by accounting for the various areas characteristics.

Table 1 presents for each dataset the root mean square error (rmse) of the mean estimator for the sampling designs described above, in addition to the simple random sampling (SRS) which is included as a comparison. The results for the stratified designs are omitted for lack of space since they were in line with the other strategies but they were never the best.

Results confirm that, as expected, when we analyse spatial-related phenomena spreading the sample over the area of interest is always convenient: the SpatLPM strategy is always better than both SRS and AuxLPM. Nonetheless, the use of the auxiliary information can improve the efficiency of the estimates, in particular if it is used to calculate the inclusion probabilities in the unequal designs.

It is important to note that, in order to evaluate when it is more or less convenient to use the auxiliary variable in addition to the geographical location, it is not enough to consider the correlation between X and Y: given the same level of correlation, estimates' efficiency depends on the proportion of co-variability that has spatial structure. If the co-variability is all defined by a spatial structure (that is, when Z_1 has $C = 50$), the SpatLPM design (with equal selection probabilities) is enough; on the other hand when part (or all) of the co-variability in not spatially related (that is, when Z_1 has $C = 30$ or $C = 0$), the additional auxiliary variable improves the estimates' efficiency, especially if used to define the unequal inclusion probabilities (UneqLPM). Moreover, UneqLPM performs better than SpatLPM even when the relation between X and Y is not linear (but still positive).

Finally, SeqUneqLPM works better than UneqLPM when the performance of the latter is worse than that of SpatLPM but it does not always manage to reach or improve the performance of the UneqLPM when this is better than the SpatLPM. These last results are very preliminary, as the experiments are still ongoing. Investigation is required on the possibility to modify the sequential procedure in order to consider more phases in which to update the inclusion probability. Moreover, experiments with more additional explanatory variables are in plan.

References

Benedetti, R., Piersimoni, F., Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*. **85**(3), pp. 439–454.

Diggle, P.J., Ribeiro, P.J. (2007). *Model-based Geostatistics*. Springer, New York.

Grafström, A., Lundström, N.L.P., Schelin, L. (1986). Spatially balanced sampling through the pivotal method. *Biometrics*, **68**, pp. 514–520.

Grafström, A., Tillé, Y. (2012). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Envirometrics*. **24**, pp. 5120–131.

Lisic, L., Cruze, N. (2016). Local Pivotal Methods for Large Surveys. In proceedings, ICES V, Geneva Switzerland 2016.

Tillé, Y. (2020). *Sampling and estimation from finite populations*. Wiley, New York.

Tillé, Y., Wilhelm, M. (2017). Probability sampling designs: Balancing and principles for choice of design. *Statistical Science*. **32**(2), pp. 176–189.

Tobler, W.R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*. **46**, pp. 234–240.

Wang, J.F., Stein, A., Gao, B.B., Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*. **2**, pp. 1–14.