Environmental and Ecological Statistics (2025) 32:1251–1267
https://doi.org/10.1007/s10651-025-00681-7

Strategies for exploiting auxiliary information and achieving spatial balance in environmental surveys

Rosa Maria Di Biase[1,2] · **Marzia Marcheselli**[1,2] · **Caterina Pisani**[1,2]

Received: 8 April 2025 / Revised: 7 August 2025 / Accepted: 20 September 2025 /
Published online: 23 October 2025

Abstract

In environmental and ecological surveys, estimating totals or functions of totals is typically approached within a design-based framework. Nowadays remote sensing technologies provide a large amount of auxiliary variables, which can be adopted at design or estimation level to improve the precision of estimators. At design level, schemes explicitly tailored to achieve spatial balance can effectively leverage these auxiliary variables. On the other hand, auxiliary variables can be adopted at estimation level both under the explicitly tailored schemes and under straightforward schemes which also ensure spatially balanced samples. An extensive simulation study compares the performances of alternative strategies exploiting auxiliary information at different levels, highlighting the challenges involved in choosing the most suitable auxiliary variables.

Keywords Auxiliary variables · Model-assisted estimators · Population of areas · Well spread samples

Handling Editor: Luiz Duczmal.

✉ Rosa Maria Di Biase
rosa.dibiase@unisi.it

Marzia Marcheselli
marzia.marcheselli@unisi.it

Caterina Pisani
caterina.pisani@unisi.it

[1] Department of Economics and Statistics, University of Siena, Piazza San Francesco 7, 53100 Siena, Italy

[2] NBFC, National Biodiversity Future Center, Piazza Marina 61, 90133 Palermo, Italy

1 Introduction

Evaluating and monitoring natural resources and biodiversity are crucial in environmental and ecological surveys and they often involve the estimation of parameters which can be expressed as totals and functions of totals. Indeed, habitat coverage, species richness and abundance are totals, and any diversity index is typically a function of the species abundance vector.

Design-based inference has become increasingly popular for estimating the total of environmental and ecological attributes, because it avoids the often unrealistic assumptions about the superpopulation probability model generating the population under study. Indeed, in the design-based framework, populations often consist of fixed sets of units scattered over the study region (e.g., the shrubs in a natural reserve, the trees in a forest) with a fixed value of an attribute of interest attached to each unit.

Enhancing the precision of total estimators heavily relies on strategically spreading the sampled units throughout the study region. This is due to two key characteristics of environmental populations: positive spatial autocorrelation and spatial heterogeneity. Positive spatial autocorrelation implies that neighboring units tend to exhibit similar values of the survey variable, reflecting the effects of shared environmental influences (Stevens and Olsen, 2004). Meanwhile, spatial heterogeneity denotes variations in the survey variable across different portions of the study region, highlighting the complexity and diversity of environmental conditions.

To address these challenges, many schemes have been proposed in literature to achieve “well spread" samples (see e.g., Stevens and Olsen, 2004; Grafström et al., 2012; Grafström and Tillé, 2013; Jauslin and Tillé, 2020) which, following Grafström (2012) and Grafström and Matei (2018) among others, are also denoted as “spatially balanced" samples. However, generally these schemes can be implemented when the unit locations are known and thus their use is precluded with environmental and ecological populations, since the list, and obviously the location, of the units are not available, especially over large areas.

On the other hand, when the study region is partitioned into a finite collection of areas, a with-list population of areal units is available with locations given by the area centroids. Populations of areal units might consist of a network of regular polygons, such as grid cells or pixels as it is common in forest inventories, or of irregular patches, such as administrative districts. In this case, the survey variable generally represents the overall amount of the attribute of interest within each area (e.g., tree biomass, dead wood) and the parameter to be estimated is the total amount of the survey variable. Obviously, the total of the attribute of interest for the population of units located on the study region coincides with the total of the survey variable of the population of areal units and estimation can be carried out selecting a probabilistic sample of areal units using a suitable sampling scheme for with-list populations.

More precisely, sampling areal units can be performed by means of the plethora of schemes explicitly tailored for selecting spatially balanced samples or by means of two schemes which can be straightforwardly adopted in environmental and ecological surveys: one-per-stratum stratified sampling and systematic sampling. Di Biase et al. (2025) compare the performance of one-per-stratum stratified sampling and systematic sampling with that of some specifically tailored spatially balanced

schemes under constant first-order inclusion probabilities, showing that no schemes outperforms the others.

An improvement in the precision of estimators can be achieved through the incorporation of inexpensive or freely available auxiliary information. Auxiliary variables like land cover classifications, topographic features, and remotely sensed imagery provide valuable additional information that can be utilized at the design level or at the estimation level. In principle, the same auxiliary information could be adopted at both design and estimation levels. However, once the information is effectively used at one level, its additional use at the other level becomes redundant and it should not provide additional benefits in terms of accuracy.

The sampling schemes explicitly tailored for achieving spatial balance can leverage auxiliary information by appropriately calibrating the first-order inclusion probabilities and/or by considering distances between area centroids not only in the geographic space, but also in the space of the auxiliary variables. Also one-per-stratum stratified sampling and systematic sampling can be implemented with first-order inclusion probabilities proportional to an auxiliary variable. However, systematic sampling is feasible only when the areal units are regular polygons grouped in blocks of equal shape and size.

Auxiliary variables can also be incorporated at the estimation level through model-assisted estimators (Breidt and Opsomer, 2017). Among the latter, starting from the generalized linear model, which underlies the generalized regression estimator, more sophisticated models have been proposed to better describe the functional relationship between the survey variable and the auxiliary variables (see e.g., Särndal et al., 1992, Montanari and Ranalli, 2005, Dagdoug et al., 2023).

Referring to populations of areal units, the aim of this paper is to explore the use of auxiliary variables in the context of spatially balanced samples. Regarding the use of auxiliary variables at the estimation level, among the wide variety of possible assisting models, the generalized regression estimator is considered, as it is attractive due to its simplicity and properties. Through an extensive simulation study, we compare the performance of alternative strategies which exploit the auxiliary information at the design or at the estimation level. The contents of the paper are organized as follows. In Section 2, notation and setting are introduced and in Section 3 a brief overview of some spatially balanced sampling schemes is given. The use of auxiliary information at design and estimation level is described in Section 4 and 5, respectively, while in Section 6 an extensive simulation study is presented. Finally, Section 7 is devoted to the discussion.

2 Notation and setting

Given a population of units scattered over the study region $\mathcal{A}$ with a fixed value of the attribute of interest for each unit, let U be the population of N areal units partitioning $\mathcal{A}$, identified by the N first integers. Notably, U is a with-list finite populations of areal units, where each unit location is given by the centroid of the area. Moreover, denote by y_j the value of the survey variable in the j-th areal unit, which is the sum of the values of the attribute of interest for the units located in the j-th areal unit.

The population total is given by

$$T = \sum_{j \in U} y_j$$

and usually constitutes the parameter to be estimated by means of a sampling strategy. In particular, let $S \subset U$ be a sample of size n, selected using a sampling design giving rise to first- and second-order inclusion probabilities π_j and π_{jh} $(h > j = 1, \ldots, N)$, respectively. Moreover, denote by $\hat{T}$ an estimator of T, that is a function of the values y_j, $j \in S$. Obviously, the aim is to choose a strategy which allows to estimate the parameter as precisely as possible.

When auxiliary variables are available, which are usually assumed to be known for each areal unit prior to sampling, their judicious use can remarkably increase the precision of the estimators. Let x_j be the vector of the values of k auxiliary variables for areal unit j, $j \in U$. As well known, auxiliary variables can be used at the design level, by calibrating first-order inclusion probabilities, or at the estimation level, by considering model-assisted estimators which leverage the relationship between the auxiliary variables and the survey variable.

3 The use of auxiliary information at design level in spatially balanced sampling

When auxiliary information is adopted at the design level, total estimation can be unbiasedly performed by means of the Horvitz-Thompson (HT) estimator

$$\widehat{T}_{HT} = \sum_{j \in S} \frac{y_j}{\pi_j}.$$

Moreover, when a fixed size sampling scheme is considered, the Sen-Yates-Grundy variance expression is

$$\mathrm{V}(\widehat{T}_{HT}) = \sum_{j \in U, h > j} (\pi_j \pi_h - \pi_{jh}) \Big(\frac{y_j}{\pi_j} - \frac{y_h}{\pi_h} \Big)^2. \tag{1}$$

If all the second-order inclusion probabilities are strictly positive, $\mathrm{V}(\widehat{T}_{HT})$ can be unbiasedly estimated by using the Sen–Yates–Grundy estimator. When some second-order inclusion probabilities are equal to zero, such as when considering designs that avoid the selection of contiguous areal units, conservative estimators can be adopted (see e.g., Wolter, 2007).

From (1), it is apparent that the precision of the HT estimator strictly depends on the first- and second-order inclusion probabilities. In particular, if a non-negative auxiliary variable is known to be approximately proportional to the survey variable, then implementing a sampling scheme with first-order inclusion probabilities proportional to that auxiliary variable will lead to small values of (1).

The challenge of constructing sampling schemes that yield designs with first-order inclusion probabilities proportional to an auxiliary variable has been a focus in the literature for a long time. Already since the early 1980s, Brewer and Hanif (1983) provided a list of 50 sampling schemes to select a sample with unequal first-order inclusion probabilities but differing in the second-order inclusion probabilities.

In environmental and ecological surveys, sampling schemes should be implemented in such a way that they not only ensure inclusion probabilities proportional to a suitable auxiliary variable but also achieve spatial balance. Indeed, neighboring areal units often exhibit similar values for the variable of interest as they interact with one another and tend to be influenced by the same natural factors (Stevens and Olsen, 2004). Grafström and Lundström (2013) make heuristic arguments regarding how spatially balanced samples can increase the precision of the HT estimator, while Grafström and Tillé (2013), exploiting the anticipated variance of the HT estimator, give a rigorous justification for using spatially balanced schemes in presence of positive autocorrelations decreasing with distance.

Many schemes explicitly tailored to achieve spatial balance allow for first-order inclusion probabilities proportional to a suitable auxiliary variable. Rather than offering a comprehensive or exhaustive review of these schemes, the following list highlights those available in open-source packages working for population and sample sizes that are sufficiently large, enabling their application by field scientists and practitioners.

Among those schemes, Generalized Random Tessellation Stratified (GRTS) sampling (Stevens and Olsen, 2004) stands out as one of the earliest proposed approaches. This scheme involves recursively dividing a grid superimposed on the study area into quadrants and assigning each quadrant a hierarchical, ordered spatial address. These spatial addresses are subsequently rearranged and projected onto a one-dimensional space, preserving spatial proximity. Finally, sampling is performed using a systematic design.

Grafström (2012) introduces Spatially Correlated Poisson Sampling (SCPS), a modification of the Correlated Poisson Sampling (CPS) method originally proposed by Bondesson and Thorburn (2008). CPS is a sequential list-based sampling scheme developed to achieve prescribed first-order inclusion probabilities. Specifically, each population unit is considered sequentially based on its position in the list and its inclusion in the sample is randomly determined. The inclusion probabilities of the remaining units are then updated using weights. SCPS incorporates weights that account for the spatial distances between units. More precisely, these weights are chosen to induce a strong negative correlation between the sample inclusion indicators of nearby units, effectively preventing the selection of units close to those already sampled.

The introduction of a distance between population units allows Grafström et al. (2012) to extend the Pivotal Method (PM) by Deville and Tillé (1998) to incorporate a spatial component, resulting in the Local Pivotal Methods (LPM). PM is a sequential sampling scheme achieving prescribed first-order inclusion probabilities where, at each step, the inclusion probabilities for two units are updated in such a way that the outcome for at least one of the units is determined. Similar to SCPS, the key concept of LPM is to induce a strong negative correlation between the sample

inclusion indicators of units that are close. Under LPM, the inclusion probabilities of two nearby units are updated at each step, and, since only nearby units compete for inclusion, LPM tends to select spatially balanced samples.

It is worth noting that distances between units can be computed not only using their spatial coordinates, but also using the values of the auxiliary variables. Thus, the sampling schemes ensure the selection of well spread samples not only in the geographic space but also in the space generated by the auxiliary variables (see e.g., Grafström et al., 2012; Grafström and Tillé, 2013).

Finally, Grafström and Tillé (2013) propose to use auxiliary variables at the design level to achieve a double property of balancing, while ensuring prescribed first-order inclusion probabilities. Indeed, the resulting sample is both spatially balanced and balanced across several auxiliary variables. Specifically, this scheme, known as Doubly Balanced Spatial Sampling (DBSS), ensures that the sample is well spread, with the HT total estimator nearly equal to the population totals for these auxiliary variables. DBSS is derived by combining a generalization of the LPM and the cube method proposed by Deville and Tillé (2004). The core concept of DBSS involves repeatedly applying the “flight phase" of the cube method to a cluster of nearby units, updating their first-order inclusion probabilities while maintaining the balancing conditions. For each cluster, the sampling outcome is determined for at least one unit, which is either included or excluded from the sample.

In the context of finite populations of areas, spatial balance can be effectively achieved using one-per-stratum stratified sampling (OPSS) and systematic sampling (SYS), both of which have a long history in statistical literature (e.g., Breidt, 1995). Under OPSS, the population of areal units is partitioned into n strata, each containing approximately the same number of contiguous areas. Then one area is selected from each stratum either randomly or with probability proportional to an appropriate auxiliary variable. When the areal units partitioning the study region are regular polygons, SYS can also be applied by grouping them into n equally shaped blocks. A polygon is selected from the first block either randomly or with probability proportional to an auxiliary variable, and the selection is then systematically repeated across the remaining blocks. Although OPSS and SYS can be implemented with inclusion probabilities proportional to an auxiliary variable, in environmental surveys they are typically applied with constant probabilities in strategies where auxiliary variables are used at estimation level.

Finally, under spatially balanced sampling, second-order inclusion probabilities may be zero or nearly zero for units that are close in distance, which prevents an unbiased estimation of the variance for the HT estimator. Therefore, to assess precision, conservative estimators (e.g., Stevens and Olsen, 2004) or ad hoc estimators (e.g., Stevens and Olsen, 2003; Grafström and Schelin, 2014; Grafström and Tillé, 2013; Fattorini, 2006; Franceschi et al., 2024) are adopted.

4 The use of auxiliary information at estimation level

The Generalized Regression Estimator (GREG) (see e.g., Särndal et al., 1992) offers a versatile framework for incorporating auxiliary information at the estimation level in order to increase precision and can be considered an essential tool in survey sampling. The core concept is that predicted values $\hat{y}_j, j \in U$, of the survey variable can be constructed for all population units by fitting an assisting model that relates the survey variable to a set of auxiliary variables. The GREG estimator of the population total is given by

$$\widehat{T}_{GREG} = \sum_{j \in U} \hat{y}_j + \sum_{j \in S} \frac{(y_j - \hat{y}_j)}{\pi_j}. \tag{2}$$

The clear motivation behind this construction is the potential to achieve a highly accurate estimate by using a well-fitting assisting model that produces minimal residuals. It is worth noting that the properties of (2) are derived on the basis of the sampling design and they hold irrespective of whether the assumptions underlying the assisting model are true or not. Thus estimation procedure is model assisted and not model dependent (Särndal et al., 1992).

The great variety of possible assisting models generates a wide family of GREG estimators of the form (2). The preference for linear regression estimators within the GREG framework is driven by their simplicity. In this setting, if $(y_j, \mathrm{x}_j)\, j \in U$ looks as if it had been generated by a linear regression model ξ, assume that $y_1, \ldots, y_N$ are realizations of independent random variables $Y_1, \ldots, Y_N$, with $E_\xi(Y_j) = \boldsymbol{\beta}'\mathrm{x}_j$ and $V_\xi(Y_j) = \sigma_j^2$, where E_ξ and V_ξ denote expected value and variance with respect to ξ and $\boldsymbol{\beta}$ and $\sigma_1^2, \ldots, \sigma_N^2$ are model parameters. Under ξ, the GREG estimator can be expressed as

$$\hat{T}_{GREG} = \hat{T}_{HT} + (\mathrm{T} - \hat{\mathrm{T}})\hat{\boldsymbol{\beta}},$$

where T is the vector of the known totals of the auxiliary variables, $\hat{\mathrm{T}}$ is the vector of the HT estimators of the auxiliary variable totals, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j \in S} \frac{\mathrm{x}_j \mathrm{x}_j^T}{\sigma_j^2 \pi_j} \right)^{-1} \left(\sum_{j \in S} \frac{\mathrm{x}_j y_j}{\sigma_j^2 \pi_j} \right).$$

The GREG estimator is approximately unbiased with respect to the sampling design and its approximate variance can be expressed as the variance of the HT estimator of the total of the regression residuals. Moreover, the GREG estimator acts as a calibration estimator ensuring the weighted totals of the auxiliary variables match known population totals. The GREG estimator encompasses several specific cases of estimators, including the ratio estimator when a single auxiliary variable is linearly related to the survey variable through the origin, simple and multiple regression esti-

mator under the omoschedasticy assumption with one or more auxiliary variables, respectively.

5 Simulation study

An extensive simulation study is conducted to evaluate the performance of various strategies that incorporate auxiliary information at either the design or estimation level. The study is based on two finite populations of spatial areas derived from the Harvard Forest tree community in Massachusetts, USA (Orwig et al., 2023). The Harvard Forest community spans a 35-hectare rectangular area (500 m $\times$ 700 m) of temperate forest. As a result, dominant tree species include white pine, red oak, red maple, and eastern hemlock. The first forest census was conducted between 2010 and 2014. During the summers of 2010 and 2011, all woody stems with a diameter at breast height of at least 1 cm were recorded, along with their spatial coordinates and other attributes, such as species and above ground biomass (AGB). Additionally, the dense shrubs in the swamp section at the center of the area were surveyed in winter (2012–2014) when the ground was frozen, facilitating more accurate data collection.

Two finite populations of areas are obtained by applying two different partitioning schemes to the region. The first population consists of $N = 560$ quadrats, each with a 25-meter side, while the second consists of $N = 3500$ quadrats, each with a 10-meter side. Two survey variables are considered: AGB and abundance. For each quadrat in both populations, the AGB value is calculated as the sum of the biomass of all trees within it. Likewise, the abundance is defined as the total number of trees within each quadrat.

Moreover, to obtain auxiliary variables, twelve wall-to-wall metrics derived from LiDAR technology (model RIEGL LD321-A40) were considered (for instrument specifications and acquisition details, see https://gliht.gsfc.nasa.gov/). Specifically, these twelve auxiliary variables were constructed by aggregating the LiDAR-derived metrics at the same spatial resolution as the two populations using the “raster" package (Hijmans, 2023) in R. Among the auxiliary variables, the one most correlated with AGB in both populations is the average tree height within the quadrat (*tree_mean*), with a correlation of 0.74 when $N = 560$ and 0.49 when $N = 3500$. Regarding abundance, the most correlated variable is the Vertical Distribution Ratio of trees (*tree_vdr*), with a correlation of 0.75 in the first population and 0.62 in the second. Figures 1 and 2 display maps of AGB, tree abundance, *tree_mean*, and *tree_vdr* for the first and second populations, respectively.

To evaluate the use of the auxiliary variables at design level, sampling is performed by means of LPM, SCPS, GRTS, and DBSS. The R package “BalancedSampling” (Grafström et al., 2022) is used to select samples by means of LPM, SCPS, and DBSS. In particular, LPM of type 1 (Grafström et al., 2012) and SCPS with maximal weights (Grafström, 2012) are performed. GRTS is implemented by means of the “spsurvey” package (Dumelle et al., 2023). For each survey variable, all sampling schemes are implemented with first-order inclusion probabilities proportional to the size of the most correlated auxiliary variable, that is *tree_mean* for AGB and *tree_vdr* for abundance. Moreover, when implementing LPM, SCPS, and DBSS, distances

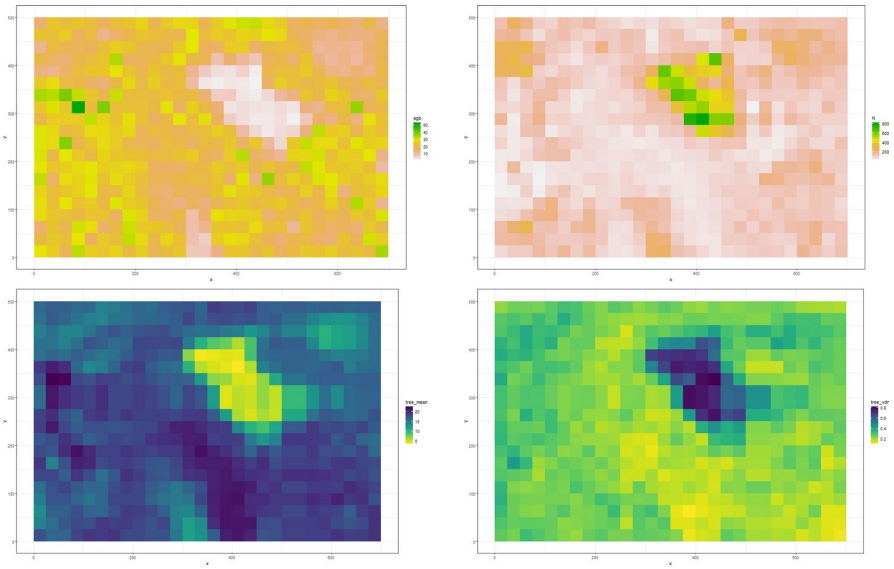

Fig. 1 Population of 560 25-m side quadrats. Each quadrat is colored according to the total above ground biomass (top left), the number of trees (top right) and the value of the LiDAR metrics *tree_mean* (bottom left) and *tree_vdr* (bottom right)

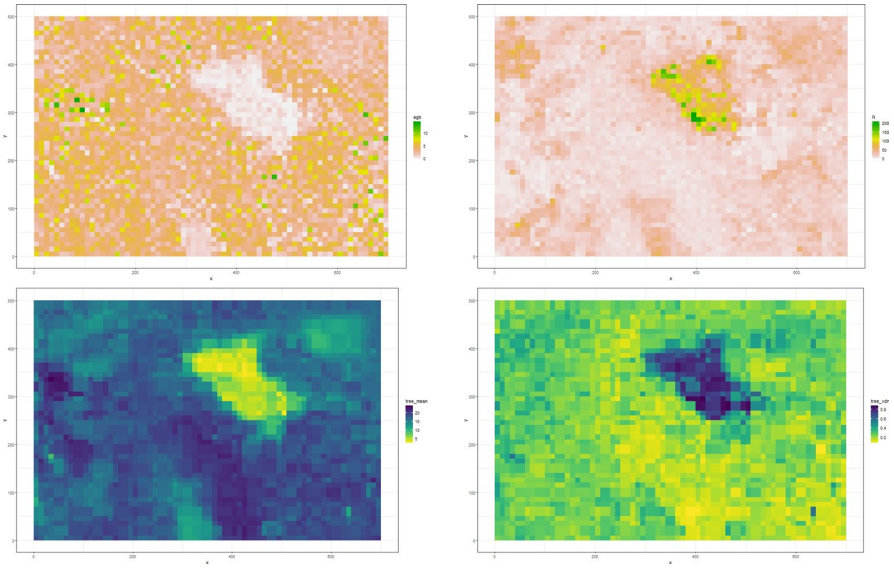

Fig. 2 Population of 3500 10-m side quadrats. Each quadrat is colored according to the total above ground biomass (top left), the number of trees (top right) and the value of the LiDAR metrics *tree_mean* (bottom left) and *tree_vdr* (bottom right)

between the centroids of the areal units are first computed using only their geographic coordinates, then using geographic coordinates and the values of the same auxiliary variable adopted to calibrate first-order inclusion probabilities and finally considering the geographic coordinates and the values of all the available auxiliary variables. In the following tables, the three different choices of variables used to calculate distances between centroids are labeled as *geo*, *geo1,* and *geoall*, respectively. DBSS is also balanced across the geographic coordinates and all the auxiliary variables.

For both populations, increasing sampling fractions of 5%, 10%, and 25%, corresponding to sample sizes of $n = 28, 56, 140$ for the first population and of $n = 175, 350, 875$ for the second population, are considered. For each combination of population, sampling scheme and sampling fraction, $R = 10000$ samples are independently selected, and the total AGB and the total abundance are estimated by means of the HT estimator. Next, the Relative Root Mean Squared Error (RRMSE) is computed as the ratio between the standard deviation, obtained by the Monte Carlo distributions of the estimates, and the parameter value.

To assess the impact of the choice of the auxiliary variable in multipurpose surveys, the procedure is repeated for each survey variable using the same sampling schemes implemented considering the auxiliary variable most correlated with the other survey variable. Specifically, *tree_vdr* is adopted to calibrate first-order inclusion probabilities and to compute distances between centroids when estimating total AGB, while *tree_mean* is adopted when estimating total abundance. The correlation coefficient of AGB with *tree_vdr* is -0.61 for $N = 560$ and -0.38 for $N = 3500$. The correlation coefficient of abundance with *tree_mean* is -0.74 for $N = 560$ and -0.75 for $N = 3500$.

Tables 1 and 2 present the RRMSE percent values for the HT estimator of the total of AGB for all sampling fractions, corresponding to the populations of $N = 560$ and $N = 3500$, respectively. Similarly, Tables 3 and 4 display the RRMSE percent values for the HT estimator of the total abundance under the same framework.

When the auxiliary information is used at the estimation level, the GREG estimator is applied using a linear assisting model. This model includes a single regressor

Table 1 RRMSE (%) of the HT estimator for total AGB using as auxiliary variable *tree_mean*, which is most correlated with AGB, or *tree_vdr*, which is most correlated with abundance, under different sampling schemes, sets of variables for computing distances, and sampling fractions *f* for the population of $N = 560$ quadrats

Scheme	*tree_mean*			*tree_vdr*		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
GRTS	3.87	2.65	1.50	7.73	5.02	2.70
SCPS *geo*	3.77	2.53	1.42	7.03	4.43	2.31
SCPS *geo1*	3.74	2.55	1.45	5.38	3.32	1.78
SCPS *geoall*	3.89	2.54	1.47	5.29	3.51	1.91
LPM *geo*	3.79	2.54	1.47	7.31	4.63	2.49
LPM *geo1*	3.81	2.57	1.47	5.79	3.58	1.95
LPM *geoall*	3.86	2.61	1.48	5.72	3.64	1.98
DBSS *geo*	3.77	2.54	1.42	4.90	3.03	1.63
DBSS *geo1*	3.84	2.60	1.46	4.87	3.11	1.70
DBSS *geoall*	3.84	2.60	1.51	4.92	3.15	1.67

Table 2 RRMSE (%) of the HT estimator for total AGB using as auxiliary variable *tree_mean*, which is most correlated with AGB, or *tree_vdr*, which is most correlated with abundance, under different sampling schemes, sets of variables for computing distances, and sampling fractions f for the population of $N = 3500$ quadrats

Scheme	*tree_mean*			*tree_vdr*		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
GRTS	3.55	2.42	1.40	5.12	3.45	2.03
SCPS *geo*	3.48	2.44	1.42	4.86	3.37	1.95
SCPS *geo1*	3.45	2.40	1.40	4.24	2.92	1.71
SCPS *geoall*	3.42	2.39	1.37	4.35	3.03	1.78
LPM *geo*	3.55	2.43	1.41	5.00	3.33	1.98
LPM *geo1*	3.42	2.37	1.39	4.41	3.01	1.73
LPM *geoall*	3.50	2.38	1.36	4.49	3.05	1.76
DBSS *geo*	3.44	2.38	1.38	4.05	2.82	1.66
DBSS *geo1*	3.44	2.38	1.37	4.07	2.79	1.67
DBSS *geoall*	3.42	2.34	1.38	4.05	2.77	1.67

Table 3 RRMSE (%) of the HT estimator for total abundance using as auxiliary variable *tree_vdr*, which is most correlated with abundance, or *tree_mean*, which is most correlated with AGB, under different sampling schemes, sets of variables for computing distances, and sampling fractions f for the population of $N = 560$ quadrats

Scheme	*tree_vdr*			*tree_mean*		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
GRTS	7.49	4.79	2.53	24.58	15.96	9.34
SCPS *geo*	6.73	4.37	2.21	22.37	13.64	7.00
SCPS *geo1*	6.94	4.47	2.28	19.00	10.17	5.38
SCPS *geoall*	6.99	4.65	2.47	19.74	11.64	5.99
LPM *geo*	6.87	4.39	2.32	23.17	14.76	7.71
LPM *geo1*	7.09	4.52	2.30	20.42	10.38	5.70
LPM *geoall*	7.12	4.71	2.45	21.00	12.31	6.05
DBSS *geo*	6.40	4.11	2.09	20.47	12.43	6.07
DBSS *geo1*	6.21	4.10	2.10	20.81	12.44	6.14
DBSS *geoall*	6.34	4.07	2.14	20.83	12.56	6.12

- specifically, the auxiliary variable most correlated with the survey variable - as well as an alternative formulation that considers all auxiliary variables as regressors. In both cases, for simplicity, the assisting regression model assumes equal variances. Sampling is performed by means of OPSS, SYS, LPM, SCPS, GRTS, and DBSS, all with equal first-order inclusion probabilities. In particular, when implementing LPM, SCPS, and DBSS, distances between the centroids of the areal units are computed using only their geographic coordinates, thus achieving spatial balance only in the geographic space, and DBSS is balanced only across the geographic coordinates. To implement OPSS and SYS, quadrats are grouped into strata/blocks of 4×5, 2×5, and 2×2 quadrats in the first population and of 10×2, 5×2, and 2×2 quadrats in the second population.

Tables 5 and 6 present the RRMSE percent values for the GREG estimator of the total AGB for each sampling fraction when using the most correlated variable as single regressor or all auxiliary variables as regressors for the population of $N = 560$

Table 4 RRMSE (%) of the HT estimator for total abundance using as auxiliary variable *tree_vdr*, which is most correlated with abundance, or *tree_mean*, which is most correlated with AGB, under different sampling schemes, sets of variables for computing distances, and sampling fractions f for the population of $N = 3500$ quadrats

Scheme	*tree_vdr*			*tree_mean*		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
GRTS	3.82	2.54	1.38	10.31	6.64	3.63
SCPS *geo*	3.58	2.35	1.25	8.24	5.18	2.84
SCPS *geo1*	3.63	2.40	1.25	7.18	4.63	2.65
SCPS *geoall*	3.73	2.47	1.34	7.67	4.97	2.75
LPM *geo*	3.69	2.39	1.27	8.92	5.61	3.07
LPM *geo1*	3.75	2.41	1.28	7.38	4.81	2.72
LPM *geoall*	3.76	2.46	1.32	7.69	4.92	2.74
DBSS *geo*	3.29	2.18	1.16	7.42	4.57	2.45
DBSS *geo1*	3.37	2.20	1.18	7.44	4.54	2.46
DBSS *geoall*	3.40	2.28	1.22	7.38	4.60	2.49

Table 5 RRMSE (%) of the GREG estimator for total AGB using either the most correlated variable, *tree_mean*, as a single regressor or all auxiliary variables as regressors, under various sampling schemes and sampling fractions f for the population of $N = 560$ quadrats

Scheme	*tree_mean*			All variables		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
OPSS	3.60	2.47	1.38	6.19	2.89	1.46
SYS	3.94	2.52	0.64	4.92	2.81	0.70
GRTS	3.82	2.52	1.44	6.75	2.99	1.49
SCPS	3.65	2.45	1.38	5.74	2.81	1.44
LPM	3.75	2.46	1.42	6.15	2.84	1.47
DBSS	3.68	2.44	1.37	5.95	2.82	1.42

Table 6 RRMSE (%) of the GREG estimator for total AGB using either the most correlated variable, *tree_mean*, as a single regressor or all auxiliary variables as regressors, under various sampling schemes and sampling fractions f for the population of $N = 3500$ quadrats

Scheme	*tree_mean*			All variables		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
OPSS	3.64	2.47	1.45	3.76	2.50	1.45
SYS	4.14	2.18	1.63	4.22	2.26	1.49
GRTS	3.60	2.46	1.43	3.73	2.49	1.44
SCPS	3.59	2.52	1.47	3.66	2.54	1.47
LPM	3.62	2.50	1.48	3.71	2.53	1.48
DBSS	3.58	2.46	1.46	3.67	2.48	1.46

and $N = 3500$, respectively. Similarly, Tables 7 and 8 display the RRMSE values for the GREG estimator of the total abundance with the same specifications.

The simulation results highlight several critical aspects regarding the choice of auxiliary variables, which heavily impacts the effectiveness of different sampling schemes in achieving estimation precision. The results clearly show that choosing an appropriate auxiliary variable at the design level is crucial for maintaining estimator

Table 7 RRMSE (%) of the GREG estimator for total abundance using either the most correlated variable, *tree_vdr*, as a single regressor or all auxiliary variables as regressors, under various sampling schemes and sampling fractions f for the population of $N = 560$ quadrats

Scheme	*tree_mean*			All variables		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
OPSS	8.54	5.59	2.93	12.07	5.97	2.94
SYS	7.97	4.63	2.28	9.87	4.86	2.78
GRTS	9.20	5.77	3.11	13.11	6.38	3.13
SCPS	8.27	5.28	2.73	11.16	5.90	2.75
LPM	8.50	5.32	2.80	11.94	5.89	2.83
DBSS	8.45	5.24	2.71	11.98	5.88	2.80

Table 8 RRMSE (%) of the GREG estimator for total abundance using either the most correlated variable, *tree_vdr*, as a single regressor or all auxiliary variables as regressors, under various sampling schemes and sampling fractions f for the population of $N = 3500$ quadrats

Scheme	*tree_mean*			All variables		
	$f = 0.05$	$f = 0.10$	$f = 0.25$	$f = 0.05$	$f = 0.10$	$f = 0.25$
OPSS	4.47	2.81	1.51	4.31	2.70	1.48
SYS	3.66	1.74	1.14	3.10	1.10	0.35
GRTS	4.51	2.98	1.60	4.32	2.81	1.52
SCPS	4.13	2.69	1.46	4.06	2.61	1.42
LPM	4.25	2.74	1.50	4.11	2.65	1.45
DBSS	4.15	2.74	1.47	4.10	2.66	1.43

precision. When the auxiliary variable strongly positively correlates with the survey variable (e.g., *tree_mean* for AGB or *tree_vdr* for abundance), the values of the RRMSEs are satisfactory and remain similar across different sampling schemes. In particular, for the smallest sampling fraction percent values of RRMSEs are always lower than 4% when AGB is considered, while for total abundance they are lower than 7.5% and than 4% for $N = 560$ and $N = 3500$, respectively. However, RRMSEs dramatically worsen when the auxiliary variable used at the design level is no strongly positively correlated, such as when estimating total AGB using *tree_vdr* and total abundance using *tree_mean*.

In contrast, when auxiliary variables are incorporated at the estimation level using the GREG estimator, the differences in performance when considering the single most correlated variable or the set of all the auxiliary variables are less pronounced, provided that the sample size is sufficiently large to allow adequate estimation of the parameters of the assisting model. Notably, using all available auxiliary variables does not markedly affect the estimation precision, suggesting that variable choice at the estimation level is more flexible than at the design level. Moreover, the performances of GREG are similar across the sampling schemes regardless of whether the most correlated auxiliary variable or the full set of auxiliary variables is used.

The tables clearly show that the RRMSEs obtained using the auxiliary variable at the estimation level are comparable to those achieved only when the appropriate variable is used at the design level and that there is no single strategy that outperforms the other for any combination of population, sample size, and parameter of interest.

Finally, not surprisingly, the simulation results show that, when the auxiliary variable is used at the estimation level, none of the considered schemes performs remarkably better than the others. In particular, the performances of OPSS and SYS are generally comparable, and sometimes even superior, to the more sophisticated spatially balanced designs. This suggests that the complexity of implementing specifically tailored spatially balanced schemes may not always be justified, especially when auxiliary information is effectively incorporated at the estimation level.

6 Discussion

The increasing availability of remotely sensed information at little or no cost, that is, data acquired from satellites and aircraft-based platforms, provides a tremendous opportunity to increase estimates precision. Indeed, remote sensing data can be used as auxiliary variables to refine the estimation strategy.

Regarding the choice between using the auxiliary variables at the design or estimation level, the latter option appears to be decidedly preferable for several reasons. When used at the design level, choosing variables from the many available, to calibrate inclusion probabilities and/or compute distances in certain spatially balanced sampling schemes, must be done before the sample is drawn, relying solely on a priori information and a poor choice can heavily deteriorate the precision of the estimators. In contrast, the choice of the variables to be used at the estimation level can be guided by analyzing the relationship between the values of the survey variable and the auxiliary variables within the sample. In this case, the sample can be considered "neutral," meaning that it can be used to estimate various population parameters. Actually, environmental surveys commonly involve multiple survey variables and, once a sample has been selected, estimators based on different sets of auxiliary variables can be considered for each parameter, allowing for the choice of the one with the best performance a posteriori. On the other hand, in multipurpose surveys, identifying an auxiliary variable at the design level that maintains a strong positive correlation with all survey variables becomes particularly challenging. Moreover, avoiding the use of the auxiliary variable at design level allow to achieve a geographically spread sample, which can be an essential requirement when the region of interest encompasses several administrative areas (e.g., administrative districts, provinces) and the number of sampled units per area needs to reflect its size.

The effectiveness of an auxiliary variable is highly dependent on its correlation with the specific variable of interest. Using an auxiliary variable indiscriminately across different estimates may lead to decreases of the precision, which are more relevant when the variable is used at the design level. Moreover, the simulation study suggests that using all the available auxiliary variables in the GREG estimator, thereby avoiding the need for variable choice, does not notably impact estimation precision when the sample size is sufficiently large to allow for an adequate estimation of the model parameters. Simulation results also show that the performances of the strategies based on OPSS or SYS and GREG estimator are generally comparable to, and sometimes even better than, those based on specifically tailored spatially balanced schemes. In addition, the simplicity of the sampling selection procedure,

which, as pointed out by Tillé and Wilhelm (2017), is one of the important aspects when choosing the sampling design, renders the use of OPSS and SYS advisable. Indeed, under these schemes, the selection process is easily understood even by field scientists, whereas it remains unclear for the more complex schemes.

Overall, the joint use of OPSS or SYS with the GREG estimator emerges as a practical and efficient strategy, particularly for applications requiring the estimation of multiple parameters. This approach allows for greater adaptability in leveraging auxiliary information while avoiding the risks associated with poor variable selection at the design level.

Acknowledgements The authors thank Lorenzo Fattorini from the University of Siena for stimulating this research and providing many suggestions.
The authors also acknowledge the support of the National Biodiversity Future Center – NBFC. Funder: Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B63C22000650007, Project title "National Biodiversity Future Center - NBFC".

Author contributions The authors contributed equally to this work.

Funding Open access funding provided by Università degli Studi di Siena within the CRUI-CARE Agreement. The study was financed by the National Biodiversity Future Center - NBFC, Project code CN_00000033 - CUP B63C22000650007.

Data availability The authors declare that the data supporting the findings of this study are available within the paper.

Code availability The code used in this article will be shared on request to the corresponding author.

Materials availability The materials used in this article will be shared on request to the corresponding author.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Ethical approval and consent to participate Not applicable.

Consent for publication The authors consent to the publication of this work.

References

Bondesson L, Thorburn D (2008) A list sequential sampling method suitable for real-time sampling. Scand J Stat 35:466–483. https://doi.org/10.1111/j.1467-9469.2008.00596.x

Breidt FJ (1995) Markov chain designs for one-per-stratum spatial sampling. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC, pp 356–361

Breidt FJ, Opsomer JD (2017) Model-assisted survey estimation with modern prediction techniques. Statist Sci 32:190–20. https://doi.org/10.1214/16-STS589

Brewer KR, Hanif M (1983) Sampling with unequal probabilities. Springer Verlag New York. https://doi.org/10.1007/978-1-4684-9407-5

Dagdoug M, Goga C, Haziza D (2023) Model-assisted estimation through random forests in finite population sampling. J Am Stat Assoc 118:1234–125. https://doi.org/10.1080/01621459.2021.1987250

Deville JC, Tillé Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85:89–10. https://doi.org/10.1093/biomet/85.1.89

Deville JC, Tillé Y (2004) Efficient balanced sampling: the cube method. Biometrika 91:893–91. https://doi.org/10.1093/biomet/91.4.893

Di Biase RM, Marcheselli M, Pisani C (2025) Achieving spatial balance in environmental surveys under constant inclusion probabilities or inclusion density functions. Environmetrics 36:e286. https://doi.org/10.1002/env.2869

Dumelle M, Kincaid T, Olsen AR, Weber M (2023) spsurvey: spatial sampling design and analysis in R. J Stat Softw 105:1–2. https://doi.org/10.18637/jss.v105.i03

Fattorini L (2006) Applying the Horvitz–Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. Biometrika. https://doi.org/10.1093/biomet/93.2.269

Franceschi S, Fattorini L, Gregoire TG (2024) Exploiting nearest-neighbour maps for estimating the variance of sample mean in equal-probability systematic sampling of spatial populations. Spat Stat 64:10086. https://doi.org/10.1016/j.spasta.2024.100865

Grafström A (2012) Spatially correlated poisson sampling. J Stat Plann Inference. https://doi.org/10.1016/j.jspi.2011.07.003

Grafström A, Lundström NL (2013) Why well spread probability samples are balanced. Open J Stat. https://doi.org/10.4236/ojs.2013.31005

Grafström A, Matei A (2018) Spatially balanced sampling of continuous populations. Scand J Stat. https://doi.org/10.1111/sjos.12322

Grafström A, Schelin L (2014) How to select representative samples. Scand J Stat. https://doi.org/10.1111/sjos.12016

Grafström A, Tillé Y (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics. https://doi.org/10.1002/env.2194

Grafström A, Lundström NL, Schelin L (2012) Spatially balanced sampling through the pivotal method. Biometrics. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Grafström A, Lisic J, Prentius W (2022) BalancedSampling: Balanced and Spatially Balanced Samplinh ttps://doi.org/10.32614/CRAN.package.BalancedSampling, https://CRAN.R-project.org/package=BalancedSampling, r package version 1.6.3

Hijmans RJ (2023) raster: Geographic Data Analysis and Modeling. https://CRAN.R-project.org/package=raster, R package version 3.6-20

Jauslin R, Tillé Y (2020) Spatial spread sampling using weakly associated vectors. J Agric Biol Environ Stat 25:431–45. https://doi.org/10.1007/s13253-020-00407-1

Montanari GE, Ranalli MG (2005) Nonparametric model calibration estimation in survey sampling. J Am Stat Assoc 100:1429–144. https://doi.org/10.1198/016214505000000141

Orwig D, Foster D, Ellison A (2023) Harvard Forest CTFS-ForestGEO mapped forest plot since 2014. Harvard Forest Data Archive: HF253 (vhttps://doi.org/10.6073/pasta/818789a882a318c1d7f3fc43a2289e12

Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer Verlag, New York

Stevens DL, Olsen AR (2003) Variance estimation for spatially balanced samples of environmental resources. Environmetrics 14:593–61. https://doi.org/10.1002/env.606

Stevens DL, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99:262–27. https://doi.org/10.1198/016214504000000250

Tillé Y, Wilhelm M (2017) Probability sampling designs: principles for choice of design and balancing. Statist Sci 32:176–18. https://doi.org/10.1214/16-STS606

Wolter KM (2007) Introduction to variance estimation. Springer Verlag New York. https://doi.org/10.1007/978-0-387-35099-8