Noname manuscript No.
(will be inserted by the editor)

Adaptive Cluster Double Sampling with post stratification with application to an epiphytic lichen community

Stefano Antonio Gattone · Paolo Giordani · Tonio Di Battista · Francesca Fortuna

Received: date / Accepted: date

Abstract The implementation of an adaptive cluster sampling design often becomes logistically challenging because variation in the final sampling effort introduces uncertainty in survey planning. To overcome this drawback, an inexpensive and easy to measure auxiliary variable could be used in a two-phase survey strategy, called adaptive cluster double sampling (Félix-Medina and Thompson, 2004). In this paper, a two-phase sampling strategy is proposed which combines the idea of adaptive cluster double sampling with the principle of post-stratification. In the first-phase an adaptive cluster sample is selected by means of an inexpensive auxiliary variable. Networks from the first phase sampling are then post-stratified according to their size. In the second-phase, the network structure is used to select a subsample of units by means of stratified random sampling. The proposed sampling strategy employs stratification without requiring an a priori delineation of the strata. Indeed, the strata sizes are estimated in the course of the two-phase sampling process. Therefore, it is suitable for situations where stratification is suspected to be efficient but strata cannot be easily delineated in advance. In this framework, a new type of estimator for the population mean which mimics the stratified sampling mean estimator and an estimator of the sampling variance are proposed. The results of a simulation study confirm, as expected, that the use of post-stratification leads to gain in precision for the estimator. The proposed sampling strategy

S.A. Gattone
Disfipeq, University of Pescara, Italy
E-mail: gattone@unich.it

P. Giordani
Difar, University of Genoa, Italy

T. Di Battista
Disfipeq, University of Pescara, Italy

F. Fortuna
Disfipeq, University of Pescara, Italy

is applied for targeting an epiphytic lichen community *Lobarion pulmonariae* in a forest area of the Northern Apennines (N-Italy), characterized by several species of conservation concern.

Keywords adaptive cluster sampling · double sampling · post-stratification · rare populations · auxiliary variable · Lobarion lichen communities

1 Introduction

It is well known that standard sampling designs are very inefficient in estimating parameters of rare and clustered populations. In response to this issue, adaptive cluster sampling designs (ACS) (Thompson, 1990) have gained popularity. The basic idea is to conduct an initial sampling phase according to a standard design and whenever the variable of interest on a selected unit satisfies a given condition, neighboring units are added to the sample and surveyed. This procedure continues until no more units are found that meet the condition. Compared to conventional sampling designs, ACS can result in higher efficiency and higher rates of encountering occupied habitat and detecting rare species (Thompson, 1990). A practical concern encountered in real application of ACS is the requirement of some prior information about the rarity and the aggregation of the population under study. Otherwise, a complete adaptive search may result to be unfeasible and the total cost of the survey may run out of control. Several remedies have been proposed in the literature to overcome this drawback (for example see Thompson (2006); Gattone et al (2016); Gattone and Di Battista (2011) and reference therein). According to Thompson and Seber (1996), the condition for adaptive designs may be based on an inexpensive and easy to measure auxiliary variable. For example, in an ecological framework, the spatial distribution of species may be approximated by some auxiliary information, such as a habitat suitability variable, which allows to delineate units with different levels of species occupancy. In particular, Félix-Medina and Thompson (2004) proposed a multi-phase variant of adaptive cluster sampling called adaptive cluster double sampling (ACDS) as a tool to control the sampling effort. The ACDS design consists in a two phase sampling where in the first phase an adaptive cluster sample is selected using the auxiliary variable. In the second phase, the network structure of ACS is used to select a subsample of units by using simple random sampling. Note that only the values of the survey variable associated with the units in the final-phase subsample are recorded. This design allows the sampler to control the number of measurements of the variable of interest. In this paper we propose to combine ACDS with the principle of double sampling for stratification. As in ACDS, an adaptive cluster sample is selected in the first phase by means of an auxiliary variable. Then, the network structure is used to form the strata. In the second-phase, the survey variable is sub-sampled in each stratum with proportional allocation. We call this method ACDS with post-stratification (ACDS-PS). We talk about post-stratification since stratification is introduced once the first-phase sampling is completed. The performance of such a design depends on the relative cost of collecting the auxiliary information with respect to the variable of interest, the extent in which the total variance is reduced by stratification and the correlation between the auxiliary and the survey variable. The novelty of this work is two-fold: the use of post-stratification which may lead to gain in precision for the estimators; a new type of estimator for the population mean which mimics the very simple estimator of stratified sampling. The paper is articulated as follows: Section 2 describes the ACDS-PS design with its mean and variance estimators. Section 3 provides a simulation study to evaluate the performance of the proposed design. In Section 4 the proposed sampling design has been used for estimating the population total of an epiphytic lichen community Lobarion pulmonariae in a forest area of the Northern Apennines (N-Italy), characterized by several species of conservation concern. The paper ends with some concluding comments in Section 5.

2 Adaptive cluster double sampling with post-stratification

Suppose to have a population $U = \{u_1, u_2, ..., u_N\}$ of N units and let y_i and x_i denote the y-value and the x-value respectively, associated with u_i, $i = 1, 2, ..., N$. In this framework, Y is the survey variable while X is a binary auxiliary variable taking two values $x_i = 1$ or $x_i = 0$ according to the presence or the absence of a given characteristic. Of interest will be estimating the population mean $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$.

According to Thompson (1990), the definition of a condition C_x together with a concept of neighbourhood give rise for each unit $u_i \in U$ to a network A_i. When X is a binary variable, the condition C_x is satisfied if $x_i = 1$. The network A_i is constituted by m_i neighboring units such that $x_j = 1$ for $j \in A_i$. The units which do not satisfy the condition C_x, *i.e.* when $x_i = 0$, represent networks of size $m_i = 1$.

Hypothetically we can divide the population into strata according to the size of the networks. For example, we may think to three strata, say L_1, L_2 and L_3 with size N_1, N_2 and N_3, respectively, with $N = N_1 + N_2 + N_3$. L_1 is constituted by units which do not meet the condition, *i.e.* $L_1 = \{u_i : x_i = 0\}$. L_2 is constituted by single units belonging to low-size networks, *i.e.* $L_2 = \{u_i : x_i = 1 \cap m_i \leq Q\}$ while L_3 contains units belonging to large-size networks, *i.e.* $L_3 = \{u_i : x_i = 1 \cap m_i > Q\}$. Q may be either fixed in advance or suitably chosen once the first phase sampling is completed.

The first-phase sample starts with the selection of a sample of n units by simple random sampling without replacement (SRSWOR). This selection results in the inclusion of the values $\{x_i : i \in U_0\}$ where U_0 denotes the set of labels selected in the first sample. The first-phase is concluded by taking an ordinary ACS sample U_1 based on the values of the auxiliary variable X. Note that the number of networks in U_1 may be less then n since more than one unit in U_0 could belong to the same network.

Once the first-phase is concluded, units in U_1 are classified (post-stratification) into strata on the basis of their network size. Let n_1 be the number of units belonging to strata L_1, n_2 the number of units in L_2 and n_3 the number of units in L_3 where $n = n_1 + n_2 + n_3$. In the second-phase, a conventional sub-sample of units from each network selected in the first-phase is taken in each stratum to observe the survey variable. Note that, for networks in stratum L_1, the second-phase sample will consist of a sub-sample of networks.

The proportions $W_h = N_h/N$ of units in the whole population belonging to each stratum h are not known but they can be estimated once U_1 has been selected with $\hat{W}_h = \frac{n_h}{n}$, $h = 1, 2, 3$. The post-stratified sampling like estimator for the population mean we propose is

$$\hat{\mu}_y = \sum_{h=1}^{H} \hat{W}_h \bar{y}_h \qquad (1)$$

where $\bar{y}_h$ is the sample mean in stratum h.

For $h = 1$, the second-phase sample is obtained by selecting a conventional sub-sample of single units (networks of size one). The stratum sample mean is given by

$$\bar{y}_1 = \frac{1}{fn_1} \sum_{i \in R_2} y_i \qquad (2)$$

where R_2 denotes the sub-sample of units selected in the second-phase and f denotes the sub-sampling fraction.

For $h = 2, 3$, the stratum sample mean is given by

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{y}_i^* \qquad (3)$$

where

$$\hat{y}_i^* = \frac{1}{fm_i} \sum_{j \in U_{2i}} y_j$$

is an estimate of the i-th network mean, y_i^*, based on a SRSWR of size fm_i. U_{2i} denotes the sub-sample of units selected in the i-th network.

In the following results, the properties of the proposed estimator are considered.

Result 1 $\hat{\mu}_y$ is an unbiased estimator of the population mean μ_y.

Proof See Appendix A.

Result 2 By using the theory of conditional moments and taking into account the different phases of sampling the design-variance of $\hat{\mu}_y$ is given by

$$v(\hat{\mu}_y) = \sum_{h=1}^{H} v(\hat{W}_h \bar{y}_h) + 2 \sum_{h=1}^{H-1} \sum_{t=h+1}^{H} cov(\hat{W}_h \bar{y}_h, \hat{W}_t \bar{y}_t). \qquad (4)$$

The variance term in equation (4), for $h = 1$, is given by

$$v(\hat{W}_1\bar{y}_1) = \frac{1}{nNf}\left[N_1 - \frac{N-n}{N-1}(1-W_1)f - W_1f\right]\sigma^2_{b_1} \\ + \bar{Y}_1^2\frac{N-n}{n(N-1)}W_1(1-W_1). \tag{5}$$

while, for $h = 2, 3$

$$v(\hat{W}_h\bar{y}_h) = \frac{1}{nN}\left[\sum_{i=1}^{N_h}\frac{\sigma^2_{w_i}}{fm_i} + \frac{(N-n)(N_h-1)}{N-1}\sigma^2_{b_h}\right] \\ + \bar{Y}_h^2\frac{N-n}{n(N-1)}W_h(1-W_h) \tag{6}$$

where $\sigma^2_{w_i} = \sum_{i\in A_i}\frac{(y_i-y_i^*)^2}{m_i-1}$ is the within-network variance of the i-th network, $\sigma^2_{b_h} = \sum_{i=1}^{N_h}\frac{(y_i^*-\bar{Y}_h)^2}{N_h-1}$ is the between-network variance in stratum h and $\bar{Y}_h$ is the mean of stratum h.

The covariance term in equation (4) is given by

$$cov(\hat{W}_h\bar{y}_h, \hat{W}_t\bar{y}_t) = -\bar{Y}_h\bar{Y}_t\frac{N-n}{n(N-1)}W_hW_t. \tag{7}$$

Proof See Appendix B.

Result 3 An estimator of $v(\hat{\mu}_y)$ is given by

$$\hat{v}(\hat{\mu}_y) = \sum_{h=1}^{H}\hat{v}(\hat{W}_h\bar{y}_h) + 2\sum_{h=1}^{H-1}\sum_{t=h+1}^{H}c\hat{o}v(\hat{W}_h\bar{y}_h, \hat{W}_t\bar{y}_t). \tag{8}$$

For $h = 1$ the variance term in (8) is given by:

$$\hat{v}(\hat{W}_1\bar{y}_1) = \frac{1}{nNf}\left[N\hat{W}_1 - \frac{N-n}{N-1}(1-\hat{W}_1)f - \hat{W}_1f\right]\hat{\sigma}^2_{b_1} \\ + \left[\bar{y}_1^2 - \hat{v}(\bar{y}_1)\right]\frac{N-n}{n(N-1)}\hat{W}_1(1-\hat{W}_1) \tag{9}$$

where

$$\hat{\sigma}^2_{b_1} = \frac{1}{fn_1-1}\sum_{i\in R_2}(y_i-\bar{y}_1)^2 \tag{10}$$

and

$$\hat{v}\left(\bar{y}_1\right) = \left\{\left[\frac{1}{fn_1} - \frac{1}{N\hat{W}_1}\right] + \frac{1-\hat{W}_1}{fn^2\left[\hat{W}_1^2 - \frac{N-n}{n(N-1)}\hat{W}_1(1-\hat{W}_1)\right]}\right\}\hat{\sigma}^2_{b_1}. \tag{11}$$

For $h = 2, 3$ the variance term in (8) is given by:

$$\begin{aligned}\hat{v}(\hat{W}_h\bar{y}_h) = \frac{1}{nN}\left[\sum_{i=1}^{n_h}\frac{\hat{\sigma}^2_{w_i}}{fm_i} + \frac{(N-n)(N\hat{W}_h-1)}{N-1}\hat{\sigma}^2_{b_h}\right] \\ + \left[\bar{y}^2_h - \hat{v}(\bar{y}_h)\right]\frac{N-n}{n(N-1)}\hat{W}_h(1-\hat{W}_h) \qquad (12)\end{aligned}$$

where

$$\hat{\sigma}^2_{w_i} = \frac{1}{fm_i - 1}\sum_{i\in U_{2i}}(y_i - \hat{y}^*_i)^2 \qquad (13)$$

$$\hat{\sigma}^2_{b_h} = \frac{1}{n_h - 1}\sum_{i\in 1}^{n_h}(\hat{y}^*_i - \bar{y}_h)^2 \qquad (14)$$

and

$$\hat{v}\left(\bar{y}_h\right) = \frac{1}{n_h N\hat{W}_h}\sum_{i=1}^{n_h}\frac{\hat{\sigma}^2_{w_i}}{fm_i} + \frac{N\hat{W}_h - n_h}{N\hat{W}_h n_h}\hat{\sigma}^2_{b_h}. \qquad (15)$$

The covariance term in (8) is given by:

$$c\hat{o}v(\hat{W}_h\bar{y}_h, \hat{W}_t\bar{y}_t) = -\bar{y}_h\bar{y}_t\frac{N-n}{n(N-1)}\hat{W}_h\hat{W}_t. \qquad (16)$$

Proof See Appendix C.

3 Simulation study

In this section a simulation study will be conducted in order to evaluate the properties of the mean estimator proposed in equation (1) together with those of the variance estimator proposed in equation (8).

The survey variable is simulated as a realization of a Poisson cluster process (Diggle, 2003) within a lattice of $60\times 60 = 3600$ squared units. The number of parents is set to 6, and they are randomly placed in the study area. A Poisson distribution with mean equal to 100 is used to assign to each parent a set of offsprings. The offsprings are located at a radial distance from the center of the parent cluster selected from an Exponential distribution with mean $r = 1.5$, and at a random angle selected from a Normal distribution with mean $\Phi = 180$ and variance equal to 60^2. This lead to a rare and clustered population with mean $\mu = 0.17$, standard deviation $\sigma = 1.53$ with a proportion of non-zero units equal to 5%.

In order to study the effect of the correlation between the survey and the auxiliary variable, two variables, X^* and Z^* are simulated using the same Poisson cluster process modifying only, for the variable Z^*, the mean of the Normal distribution set to $\Phi = 0$. In this way, X^* and Y turn out to have a high level of correlation ($\rho_{X^*,Y} = 0.95$). A lower level of correlation is, instead, observed between Z^* and Y ($\rho_{Z^*,Y} = 0.75$). In order to implement the proposed ACDS-PS design, X^* and Z^* are transformed into two binary

Table 1 Expected sampling effort of ACDS-PS and initial sample of ACS

n_1	ACDS-PS		ACS	
	$E(v_{aux})$	$E(v_y)$	c_2	n_{ACS}
30	75	27	5	18
			20	13
60	140	50	5	34
			20	25

variables X and Z. For example, X was defined as $X = 1$ if $X^* > 0$ and $X = 0$ otherwise. For the couple X, Y, 2.6% of the units have both X and Y greater than zero. For the couple Z, Y, the percentage drops to 1.7%.
$M = 15000$ Monte Carlo samples are simulated from the following designs: ACDS-PS, ACDS and ACS. The estimators used in each design are $\hat{\mu}_y$ for ACDS-PS, an Horvitz-Thompson type estimator [equation (2) of Félix-Medina and Thompson (2004)] for ACDS and the Horvitz-Thompson estimator for ACS.
Both ACDS-PS and ACDS are carried out with initial samples of size $n = 30, 60$ selected by SRSWOR. Whenever the auxiliary variable is equal to one, the adaptive search is conducted into a neighborhood consisting into the four plots sharing a common boundary line. The two-phase adaptive designs differ in the second phase sampling. Under ACDS, the networks are sub-sampled by SRSWOR, whereas, under ACDS-PS, networks are sub-sampled in the strata formed after the first phase. The ACDS-PS design is implemented by setting the number of strata to $H = 3$ and the strata are identified by setting Q equal to the median of the sizes of the networks selected in the first phase. The sampling fraction at the second phase was set to $f = 0.5$ in all the strata.
The three designs are compared under the same total expected cost. The cost function is defined as $C_T = c_1 v_{aux} + c_2 v_y$, where c_1, c_2 and and v_{aux} and v_y are the per element costs and the sampling effort of the first and the second-phase sampling, respectively. In determining v_{aux}, the edge units selected in the first adaptive phase are also counted. Two values of c_2 are considered: $c_2 = 5, 20$. In this way we are considering situations in which observing the survey variable is five times or twenty times more expensive than observing the auxiliary variable. Once C_T is computed for ACDS-PS, the initial sample of the ACS design (n_{ACS}) is set so to have the same expected total cost $E(C_T)$. The expected sampling effort of ACDS-PS together with the size of the initial sample of the ACS design are reported in Table 1. Sampling effort for ACDS is not reported since it is equivalent to ACDS-PS.

The performance of each estimator is evaluated by its relative efficiency with respect to the sample mean $\bar{y}$ under SRSWOR. In particular, the relative efficiency of an estimator $\hat{\mu}$ is computed as follows:

$$Re(\hat{\mu}) = \frac{\sigma_{\bar{y}}}{rMSE(\hat{\mu})} \qquad (17)$$

Table 2 Relative efficiency of the adaptive designs

Population	c_2	$ACDS-PS$		$ACDS$		ACS	
		$n=30$	$n=60$	$n=30$	$n=60$	$n=30$	$n=60$
x,y	5	1.62	1.64	1.21	1.24	1.23	1.25
	20	1.89	1.93	1.44	1.46	1.21	1.24
z,y	5	1.35	1.39	1.21	1.22	1.23	1.25
	20	1.59	1.64	1.43	1.43	1.21	1.24

Table 3 Relative bias and relative mean squared error of the ACDS-PS variance estimator

Population	RB		$RMSE$	
	$n=30$	$n=60$	$n=30$	$n=60$
x,y	-0.025	-0.015	0.158	0.083
z,y	-0.019	0.004	0.507	0.236

where $\sigma_{\bar{y}} = \sqrt{\frac{N-n}{Nn}}\sigma$ with $n = \frac{C_T}{c_2}$ and $rMSE(\hat{\mu}) = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(\hat{\mu}_j - \mu)^2}$.

Results for both populations are listed in Table 2 for different initial sample sizes, n, and different marginal costs of measuring the auxiliary variable, c_2. As expected, with rare and clustered populations all the adaptive designs are more efficient than SRSWOR. Furthermore, both ACDS and ACDS-PS improve their efficiency from $c_2 = 5$ to $c_2 = 20$, *i.e.* in presence of an increase of the relative cost of measuring the response variable with respect to the auxiliary variable. ACDS shows better results than ACS when $c_2 = 20$ and it performs as ACS when $c_2 = 5$. On the other hand, ACDS-PS results to be the best design reporting more pronounced gains in relative efficiency when the auxiliary and the response variables have higher correlation. The initial sample size n seems to not affect the pattern of the performances of the designs considered in the simulation.
In Table 3 the performance of the variance estimator (8) is evaluated by its relative bias and its relative root mean squared error (rMSE over the standard deviation). Results show that the approximation provided for the variance is very accurate and nearly unbiased.

4 Application

The proposed ACDS-PS design is used for targeting a rare and clustered epiphytic lichen community in a forest area of the Northern Apennines (N-Italy), characterized by several species of of conservation concern (Nascimbene et al, 2013). In particular, we refer to the species of Lobarion communities which are considered an important indicator of sustainable forest management (Campbell and Fredeen, 2004; Nascimbene et al, 2010) because they are strictly dependent on forest structure and dynamics (Nascimbene et al, 2010; Giordani et al, 2015). In this study, the following species of Lobarion communities are targeted:

- *Collema nigrescens* (Huds.) DC.
- *Fuscopannaria ignobilis* (Anzi) P.M. Jϕrg.
- *Leptogium burnetiae* C.W. Dodge
- *Lobaria pulmonaria* (L.) Hoffm.
- *Lobarina scrobiculata* (Scop.) Nyl.
- *Nephroma laevigatum* Ach.
- *Nephroma resupinatum* (L.) Ach.
- *Nevesia sampaiana* (Tav.) P.M. Jϕrg., L. Lindblom, Wedin & S. Ekman
- *Pannaria conoplea* (Ach.) Bory
- *Parmeliella testacea* P.M. Jϕrg.
- *Parmeliella triptophylla* (Ach.) Müll. Arg.
- *Pectenia plumbea* (Lightf.) P.M. Jϕrg., L. Lindblom, Wedin & S. Ekman
- *Peltigera collina* (Ach.) Schrad.
- *Ricasolia amplissima* (Scop.) De Not. (chloromorph and cyanomorph).

These species have been recorded in a survey area of c.a. 2.2 Km^2 in Lame Forest (Val d'Aveto, Ligurian Apennines, NW-Italy). The study area was divided into 5456 plots of size 20 × 20 m (white plots in Fig. 1). The area has a considerable naturalistic relevance and is characterized by mixed forest formations (chestnut, beech, pine forests), with different management, interspersed with grasslands and shrubs. Although the Lame Forest is an optimal site for species of conservation interest (Nascimbene et al, 2013), the lichen species under consideration are rare and present a clustered distribution on both large and local scales. In particular, the targeted lichen species are often associated to particular habitats, such as chestnut, or to structural characteristics of the same, for example chestnut orchards with very large trees or

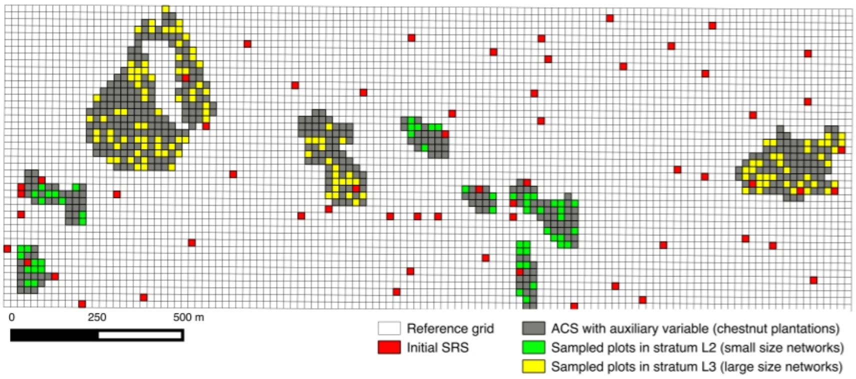
0
250
500 m
□ Reference grid
🟥 Initial SRS
■ ACS with auxiliary variable (chestnut plantations)
🟩 Sampled plots in stratum L2 (small size networks)
🟨 Sampled plots in stratum L3 (large size networks)

Fig. 1 Sampled plots in Lame Forest area.

isolated trees (Matteucci et al, 2012). Due to the biological characteristics of these organisms, the ACDS-PS design may be useful to improve the detection probabilities for rare species by achieving a cost-effective sampling effort.

In the first phase, an ACS sample is selected with an initial SRSWOR sample of size $n = 60$ units (red plots in Fig. 1). The auxiliary information used in this first phase is a binary variable corresponding to the presence-absence of Chestnut plantations coverage. The resulting ACS sample is constituted by $K = 56$ distinct networks (gray plots in Fig. 1). The latter are classified (post-stratification) into three strata (L_1, L_2, L_3) on the basis of their size, by setting $Q = 60$. The strata represent networks of size one, small and large sizes, respectively.

In the second phase, a stratified random sampling with proportional allocation is taken in each stratum to observe the survey variable y, the total number of trees colonized by at least one of the target species. For each strata, the sampling fraction is set to 0.3. The ACDS-PS sample is composed by $n = 231$ plots of which 47 networks of size one in L_1, that is units which did not meet the condition; 6 small size networks (size 26, 38, 29, 16, 25, 51) in strata L_2 (green plots in Fig. 1) and 3 large size networks (size 236, 82, 115) in strata L_3 (yellow plots in Fig. 1). In the whole survey area, a total of 2944 trees (standard error estimate equal to 515) were estimated to host at least one Lobarion species. Apparently, this is enough to ensure the conservation of Lobarion community in the forest, against the pressure of large scale disturbance factors, such as climate change, which is expected to halve by 2050, the presence of Lobarion species in Italian sites where the taxa currently occur (Nascimbene et al, 2016). Despite the considerable total number of colonized trees in the survey area, at plot level, the average number of trees with at least one target species, was very low with $\hat{\mu}_y = 0.5396$ and $\hat{v}_{\hat{\mu}_y} = 0.0943$. This may lead to fine scale extinction events, which in turn, might affect the effectiveness of dispersal and establishment of new propagules throughout the forest, interrupting the entire population dynamic.

The choice of the ACDS-PS design is driven by the relative cost of phase 1 to phase 2 samples, that is approximately equal to 1/50. Indeed, the evaluation of the auxiliary variable required about 1 hour for 500 plots compared to ca. 30 minutes per plot for measuring the survey variable in the field. Moreover, the time needed for fieldwork significantly decreased for those plots which are included in large networks (ca. 15 plots per hour) with respect to scattered plots of stratum L_1 and L_2 (less than 10 plots per hour). A possible applicative constraint of the ACDS-PS lays in its dependence on the cluster distribution of the response variable: the more the target species is locally abundant and clustered within the survey area, the more effective is the design. In the case of our application, many target species resulted even more rare and scattered than expected. Indeed, even though the auxiliary variable (occurrence of chestnut plantations) is highly correlated with the response variable, it is not able to discern other micro environmental characteristics of the habitat, which decisively drive the occurrence of Lobarion species (Nascimbene et al, 2013). Thought that these latter factors are hardly cognizable a priori, the ACSDS-

PS produced a cost-effective trade-off between time needed for fieldwork and reliability of the results. Our findings may contribute to enhance the conservation of Lobarion species at local scale. Indeed, notwithstanding these practical limitations, the ACDS-PS design was able to detect several large clusters of rare species, getting a fine-scale feature of their distribution even within each sampled network.

5 Conclusions

To address the need for a successful environmental management system, new designs for sample surveys have been developed in the last decades. The aim is to provide targeted information at the desired spatial spread combining two main requirements: to obtain information in a statistically valid way and to efficiently use the available resources. This work is about a new sampling design applied for targeting rare and sparsely distributed populations. A two-phase strategy, which suitably merge the information available from an auxiliary variable with the one provided by the survey variable, is proposed. The first-phase sample is selected using an ACS design. This shall be done using an inexpensive auxiliary variable. Networks from the first phase sampling are then post-stratified according to their size. In the second-phase, subsample of units are selected with stratified random sampling. The values of the survey variable are measured only in the units of the second, smaller sample. Note that the proposed sampling strategy employs stratification without requiring an a priori delineation of the strata. Indeed, the strata sizes are estimated in the course of the two-phase sampling process. Therefore, it is suitable for situations where stratification is suspected to be efficient but strata cannot be easily delineated in advance. The performance of such a design will depend on the extent to which the total variance is reduced by stratification. We would like to stress how the proposed design is different from some variants of ACS existing in the literature which also contain the idea of stratification and the idea of a two-phase strategy. The idea of stratification has already been introduced in ACS (Thompson, 1991) where an initial sample of units is selected from a population using stratified random sampling. Then in each strata, units are added adaptively. The proposed two-phase strategy is also different from the Two-Stage ACS (Salehi and Seber, 1997) where after taking a SRSWOR of primary units, a subsample of secondary units is taken by means of ACS. The ACDS-PS design is different since an ACS sample is selected in the first-phase by means of an auxiliary variable and the network structure is used to select in the second-phase a subsample of units by means of stratified random sampling. In this study we consider only the case where the auxiliary information is a binary variable. One of the possible developments of the current work could be dealing with a continuous auxiliary variable. This will require, for instance, the use of a ratio-type or regression-type estimator. Also, defining an optimal criterion for selecting the stratum boundaries and the sampling fraction in each stratum to reduce the variance of the proposed estimator would be a fruitful area of future research.

References

Campbell J, Fredeen A (2004) Lobaria pulmonaria abundance as an indicator of macrolichen diversity in interior Cedar-Hemlock forests of east-central British Columbia. Canadian Journal of Botany 82:970–982

Diggle P (2003) Statistical analysis of spatial point patterns. Edward Arnold, London, 2nd edition

F$\acute{e}$lix-Medina M, Thompson S (2004) Adaptive cluster double sampling. Biometrika 91:877–891

Gattone S, Di Battista T (2011) Adaptive cluster sampling with a data driven stopping rule. Statistical Methods and Applications 20:1–21

Gattone S, Mohamed E, Di Battista T (2016) Adaptive cluster sampling with clusters selected without replacement and stopping rule. Environmental and Ecological Statistics 23:453–468

Giordani P, Benesperi R, Mariotti M (2015) Local dispersal dynamics determine the occupied niche of the red-listed lichen Seirophora villosa (Ach.) Fr$\ddot{o}$d$\acute{e}$n in a Mediterranean Juniperus shrubland. Fungal Ecology 13:77–82

Matteucci E, Benesperi R, Giordani P, Piervittori R, Isocrono D (2012) Epiphytic lichen communities in chestnut stands in Central-North Italy. Biologia 67:61–70

Nascimbene J, Brunialti G, Ravera S, Frati L, Caniglia G (2010) Testing Lobaria Pulmonaria (L.) Hoffm. as an indicator of lichen conservation importance of Italian forests. Ecological Indicators 10:353–360

Nascimbene J, Benesperi R, Brunialti G, Catalano I, Vedove M, Grillo M, Isocrono D, Matteucci E, Potenza G, Puntillo D, Puntillo M, Ravera S, Rizzi G, Giordani P (2013) Patterns and drivers of β-diversity and similarity of *Lobaria pulmonaria* communities in Italian forests. Journal of Ecology 101:493–505

Nascimbene J, Casazza G, Benesperi R, Catalano I, Cataldo D, Grillo M, Isocrono D, Matteucci E, Ongaro S, Potenza G, Puntillo D, Ravera S, Zedda L, Giordani P (2016) Climate change fosters the decline of epiphytic *Lobaria* species in Italy. Biological Conservation 201:377–384

Salehi M, Seber G (1997) Two-stage adaptive cluster sampling. Biometrics 53:959–970

Stephan F (1945) The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. Annals of Mathematical Statistics 16:50–61

Thompson S (1990) Adaptive cluster sampling. Journal of the American Statistical Association 85:1050–9

Thompson S (1991) Stratified adaptive cluster sampling. Biometrika 78:389–397

Thompson S (2006) Adaptive web sampling. Biometrics 62:1224–1234

Thompson S, Seber G (1996) Adaptive Sampling. Wiley, New York

A Unbiasedness of $\hat{\mu}_y$

We have

$$E\left(\hat{W}_h \bar{y}_h\right) = E\left[E\left(\hat{W}_h \bar{y}_h \mid \hat{W}_h\right)\right] = E\left[\hat{W}_h E\left(\bar{y}_h \mid \hat{W}_h\right)\right]. \tag{18}$$

For samples in which $\hat{W}_h$ is fixed, in every stratum the sample estimate $\bar{y}_h$ is unbiased, *i.e.* $E\left(\bar{y}_h \mid \hat{W}_h\right) = \bar{Y}_h$. Then,

$$E\left[\hat{W}_h E\left(\bar{y}_h \mid \hat{W}_h\right)\right] = E\left(\hat{W}_h\right)\bar{Y}_h = W_h \bar{Y}_h. \tag{19}$$

Finally, $E\left(\hat{\mu}_y\right) = E\left(\sum_{h=1}^{H} \hat{W}_h \bar{y}_h\right) = \sum_{h=1}^{H} W_h \bar{Y}_h = \mu_y$.

B Design variance of $\hat{\mu}_h$

The variance term in equation (4) is equal to

$$\begin{aligned} v\left(\hat{W}_h \bar{y}_h\right) &= E\left[v\left(\hat{W}_h \bar{y}_h \mid \hat{W}_h\right)\right] + v\left[E\left(\hat{W}_h \bar{y}_h \mid \hat{W}_h\right)\right] \\ &= E\left[\hat{W}_h^2 v\left(\bar{y}_h \mid \hat{W}_h\right)\right] + \bar{Y}_h^2 v\left(\hat{W}_h\right) \\ &= E\left[\hat{W}_h^2 v\left(\bar{y}_h \mid \hat{W}_h\right)\right] + \bar{Y}_h^2 \frac{N-n}{(N-1)n} W_h(1-W_h). \end{aligned} \tag{20}$$

For $h = 2, 3$, the conditional variance in the first term of equation (20) can be obtained using the results of Thompson (1990) for the Hansen-Hurwitz estimator and considering the variability in the two phases of sampling.

$$v\left(\bar{y}_h \mid \hat{W}_h\right) = \frac{1}{n_h N_h} \sum_{i=1}^{N_h} \frac{\sigma_{w_i}^2}{m_i} + \frac{N_h - n_h}{N_h n_h} \sigma_{b_h}^2. \tag{21}$$

The second term of equation (21) is the variance of the Hansen-Hurwitz estimator in ordinary ACS while the first term represents the increase in variance that arises because of the subsampling of networks in the second phase.

Substituting equation (21) in the first term of (20), after some algebraic manipulations and noting that $E(n_h) = nW_h$ and $E(n_h^2) = \frac{N-n}{N-1} nW_h(1-W_h) + n^2 W_h^2$ we obtain

$$E\left[\hat{W}_h^2 v\left(\bar{y}_h \mid \hat{W}_h\right)\right] = \frac{1}{nN}\left[\sum_{i=1}^{N_h} \frac{\sigma_{w_i}^2}{m_i} + \frac{(N-n)(N_h-1)}{N-1}\sigma_{b_h}^2\right]. \tag{22}$$

Substituting (22) in equation (20) we obtain equation (6) of section 2.

In the first strata, $h = 1$, $\sigma_{w_i}^2 = 0$ since the networks have size one. Then equation (21) changes to

$$v\left(\bar{y}_1 \mid \hat{W}_1\right) = \frac{N_1 - n_1 f}{N_1 n_1 f} \sigma_{b_1}^2 \tag{23}$$

Substituting equation (23) in the first term of (20), after some algebraic manipulations we obtain

$$E\left[\hat{W}_1^2 v\left(\bar{y}_1 \mid \hat{W}_1\right)\right] = \frac{1}{nNf}\left[N_1 - \frac{N-n}{N-1}(1-W_1)f - W_1 f\right]\sigma_{b_1}^2. \tag{24}$$

Substituting (24) in equation (20) we obtain equation (5) of section 2.

Finally, applying the law of total covariance allows us to express the covariance term in equation (4) as follows:

$$cov\left(\hat{W}_h\bar{y}_h,\hat{W}_t\bar{y}_t\right)=E\left[cov\left(\hat{W}_h\bar{y}_h,\hat{W}_t\bar{y}_t\mid\hat{W}_h,\hat{W}_t\right)\right]+cov\left[E\left(\hat{W}_h\bar{y}_h\mid\hat{W}_h\right),E\left(\hat{W}_t\bar{y}_t\mid\hat{W}_t\right)\right].\tag{25}$$

The first term is clearly zero and noting that $cov\left(\hat{W}_h,\hat{W}_t\right)=-\frac{N-n}{(N-1)n}W_hW_t$ we have

$$cov(\hat{W}_h\bar{y}_h,\hat{W}_t\bar{y}_t)=-\bar{Y}_h\bar{Y}_t\frac{N-n}{(N-1)n}W_hW_t.\tag{26}$$

C Estimator of the variance of $\hat{\mu}_h$

For $h=1$, the variance term in (5) is estimated by equation (9). The first part is obtained by substituting to W_1 and $\sigma^2_{b_1}$ the estimators $\hat{W}_1=\frac{n_1}{n}$ and $\hat{\sigma}^2_{b_1}$. The second part is obtained by substituting to $\bar{Y}_1^2$ its estimate $\bar{y}_1^2-\hat{v}\left(\bar{y}_1\right)$.

In order to compute the variance of $\bar{y}_1=\frac{1}{fn_1}\sum_{i\in R_2}y_i$ we have to consider that the number of units n_1 in the sample belonging to stratum L_1 will vary from sample to sample. We have that $v\left(\bar{y}_1\right)=E\left[v\left(\bar{y}_1\mid n_1\right)\right]+v\left[E\left(\bar{y}_1\mid n_1\right)\right]$ where the first term is the variance of $\bar{y}_1$ for samples in which n_1 is fixed while the second term is zero. Thus

$$v\left(\bar{y}_1\right)=E\left[\frac{N_1-fn_1}{N_1fn_1}\sigma^2_{b_1}\right]=E\left[\frac{1}{fn_1}-\frac{1}{N_1}\right]\sigma^2_{b_1}.\tag{27}$$

From Stephan (1945) an approximation of the expected value $E\left(\frac{1}{n_1}\right)$ is given by

$$E\left(\frac{1}{n_1}\right)\approx=\frac{1}{nW_1}+\frac{1-W_1}{n^2W_1^2}.\tag{28}$$

By substituting the above result in equation (27) we end up with

$$v\left(\bar{y}_1\right)=\left[\frac{1}{fn_1}-\frac{1}{N_1}+\frac{1-W_1}{fn^2W_1^2}\right]\sigma^2_{b_1}.\tag{29}$$

The estimator $\hat{v}\left(\bar{y}_1\right)$ of $v\left(\bar{y}_1\right)$ provided in equation (11) is obtained by estimating N_1 with $N\hat{W}_1$, W_1 with $\hat{W}_1$, W_1^2 with $\hat{W}_1^2-\frac{N-n}{n(N-1)}\hat{W}_1(1-\hat{W}_1)$ and $\sigma^2_{b_1}$ with $\hat{\sigma}^2_{b_1}$.

Similarly, for $h=2,3$, we can obtain the estimator (12) of the variance term in (6).