Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

A sampling strategy for assessing habitat coverage at a broad spatial scale

Lorenzo Fattorini [a,1], Marco Cervellini [b,c,1], Sara Franceschi [a,*], Michele Di Musciano [b,d], Piero Zannini [b,e], Alessandro Chiarucci [b,e]

[a] *Department of Economics and Statistics, University of Siena, Siena, Italy*
[b] *BIOME lab, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum, University of Bologna, Bologna, Italy*
[c] *School of Biosciences and Veterinary Medicine, Plant Diversity and Ecosystems Management Unit, University of Camerino, Camerino, Italy*
[d] *Department of Life, Health & Environmental Sciences, University of L'Aquila, Coppito, L'Aquila, Italy*
[e] *Inter University Centre PlantDATA, Department of Biological, Geological and Environmental Sciences, Alma Mater Studiorum, University of Bologna, Bologna, Italy*

ARTICLE INFO

Keywords:
Design-based inference
Italy
Monitoring
Precision estimation
Two-phase sampling
Two-stage sampling
Simulation study

ABSTRACT

The quantitative assessment of habitat conservation status is a major task for European Union member states in compliance with Council Directive 92/43. One goal of the European 2030 Biodiversity Strategy is the effective management of habitats that show declining trends. While various approaches have been adopted for national assessments, there is no consensus on how to achieve common statistically sound estimates of the criteria indicated by the EU Directive for the evaluation of the status and trend of habitat types. Here, we present an adaptive monitoring approach based on a two-phase sampling scheme to estimate the coverage of EU terrestrial habitat types, which is one of the four criteria indicated by the Habitats Directive. We used 9 habitats distributed among different EU member states choosing Italy as a case study. The development of the methodological approach is described, and a simulation study was performed to check the precision of the coverage estimators accounting for the lack of sampled data (nonresponse treatment), subregions and sustainable sampling effort. We found that our two-phase sampling approach has the potential to increase precision in estimating the coverage of habitat types (approximated at 1 ha cell size) with respect to the precision achieved by simple random sampling without replacement, which is the simplest sampling approach. Adopting a small sampling fraction (⩽0.04%) of the survey area, the relative standard errors ranged from 7 to 15% for common habitats whose presence is strongly correlated with the habitat suitability scores furnished by an expert team. In the challenging context of a "mandated" monitoring type, our approach provides sound statistical estimates of habitat coverage with the possibility of applying a standardised and transferable sampling scheme that is easily repeatable over time.

1. Introduction

The decline of habitat structure and functioning is becoming increasingly relevant worldwide, not only because of the consequent dramatic biodiversity loss but also due to the quantitative and qualitative decrease in various ecosystem services (Martinez-Harms et al., 2015; Mulder et al., 2015; Keyes et al., 2021). The need to know how habitats are changing is becoming a pivotal challenge in applied ecology, and science-based information provides solid strategic information for global decision-making. In this direction, long-term ecological research is necessary to fulfil ecological and social goals, and it should be founded on well-designed long-term adaptive (sensu Lindenmayer and Likens, 2009, 2018; Lindenmayer et al., 2020) monitoring approaches (Smith and Gray, 2021; Cowles et al., 2021).

In Europe, the conservation of EU habitat types within and outside of the world's largest coordinated network of protected areas (Natura 2000 Network, hereafter N2K; European Commission, 2021) is one of the main targets of EU Directive 92/43. Furthermore, the EU Biodiversity Strategy 2030 targets the enforcement of the Habitats Directive (HD) by enlarging the N2K and by improving the effective management of sites and habitats that show declining trends (European Commission, 2020). Under this framework, it is essential to achieve quantitative and affordable measures of the status and trends concerning habitat area and quality.

* Corresponding author.
E-mail addresses: lorenzo.fattorini@unisi.it (L. Fattorini), marcocervellini@gmail.com (M. Cervellini), sara.franceschi@unisi.it (S. Franceschi), michele.dimusciano@univaq.it (M. Di Musciano), piero.zannini2@unibo.it (P. Zannini), alessandro.chiarucci@unibo.it (A. Chiarucci).

[1] Equally contributing authors.

https://doi.org/10.1016/j.ecolind.2022.109352
Received 20 February 2022; Received in revised form 11 August 2022; Accepted 21 August 2022
Available online 2 September 2022

The HD describes a “habitat” as an area of a species’ occurrence defined by geographic, abiotic, and biotic features. The ecological literature provides a variety of alternative habitat definitions contributing to a detailed qualitative description of the autecological species-related habitat concept, while an operational and standardized quantitative characterization of this concept is still lacking (e.g., SYapp, 1922; Hall et al., 1997; Davies et al., 2004; Mitchell, 2005; Drakou et al., 2011; Cervellini et al., 2021). In this context, Fahrig, 2013 recently proposed the “habitat patch concept”. This concept assumes that habitat patch boundaries contain and delimit biological populations and communities and raises the problem of how to delineate and measure ecologically relevant habitat patches. Although new approaches to measure and model the terrestrial area of a habitat are discussed (e.g., area of occupancy “AOO” and area of habitat “AOH”, see Álvarez-Martínez et al., 2018; Brooks et al., 2019), complete habitat type mapping is still a challenge, particularly when the habitat classification and the related conservation goals differ among organizations or regulations (e.g., IUCN and HD). In summary, the debate between categorical and dynamic habitat maps is ongoing (discrete classes of habitat vs. continuous values, see Coops and Wulder, 2019), and a univocal and operational definition of the parameter “area of habitat” is still lacking.

Formally, the evaluation of the conservation status of EU habitat types is based on four criteria: “range”, “area”, “habitat quality” and “pressures and threats”, but a standardized approach for ecological monitoring at the EU scale is still lacking (Ellwanger et al., 2018; Lengyel et al., 2018; Delbosc et al., 2021). This gap is quite understandable in the complex context of habitat recognition and mapping. Despite methodological problems, monitoring the conservation status of habitats listed in Annex I is mandatory according to Article 11 of EU Directive 92/43. Each member state (MS) must submit a national report to the European Commission every six years on the implemented measures and their effectiveness (Art. 17 HD) based on monitoring results. Most EU countries are producing six-year reports based on expert-based assessments or supposed complete censuses. For instance, in Italy, during the past four reporting cycles (1994–2018), the Institute for Environmental Protection and Research (ISPRA) provided the European Commission with a habitat conservation status assessment for both national and biogeographical regions by merging the data independently gathered by the 21 Italian regions and autonomous provinces. The habitat monitoring actions performed by these local public agencies or institutions were based on standardized guidelines concerning “how” to survey in the field (Angelini et al., 2016) but without indications about “where” (i.e., sampling scheme) and “how much” to survey (i.e., sampling effort). These omissions made it extremely difficult to fully merge the data and perform statistical inferences on countrywide habitat population and each biogeographical region and to quantify and detect changes or trends in the targeted criteria between the different reporting cycles. Ellwanger et al., 2018 showed that none of the surveyed MSs made a theoretical statement on the statistical strength of the adopted monitoring approaches, highlighting the need for better sampling and assessment approaches.

Developing an adaptive and statistically sound long-term monitoring plan is becoming pivotal to establish how data should be collected and to produce standardized, reliable estimates for given parameters (e.g., area). Data from such a sample survey should be (a) representative of the population under investigation and (b) information-rich to reduce uncertainty about inferences (Foster, 2020). Achieving these outcomes at the continental scale becomes particularly complex in a “mandated” monitoring plan (Lindenmayer and Likens, 2010) as that imposed by the HD. In this context, it is crucial to provide a standardized sampling strategy (Delbosc et al., 2021) to guarantee the following three properties are considered generally relevant in determining the scientific quality of biodiversity monitoring (Lengyel et al., 2018): (i) a sound and feasible sampling scheme, (ii) a good trade-off between sampling effort and the precision of the resulting estimators, and (iii) appropriate statistical analysis to detect changes or trends.

Here, we develop a two-phase sampling strategy to estimate quantities approximating the area of terrestrial habitats, which is one of the four criteria indicated by the HD for evaluating the conservation status of EU habitat types. After conceptual analysis and methodological development of the strategy, a simulation study was performed to check and compare the precision of the proposed estimators for nine selected habitats, seven of which were distributed among different EU countries (EIONET, 2022) under several potential and real constraints (e.g., critical aspect emerged during the previous four reporting cycles) that may arise during the surveys (e.g., presence of auxiliary information, nonresponses). This approach was developed for the territory of a single country, namely, Italy, but can be rescaled to the territory of any other country or to the entire European Union.

2. Materials and Methods

2.1. Study region

The study region was the surface area within the administrative borders of the Italian state. It spans $301,328.46km^2$ and is covered by $3,491$ quadrats of $10km \times 10km$ of the grid used for reporting HD data by European member states (European Enviroment Agency, 2013; Cervellini et al., 2020). Therefore, the total area of the grid overlapping the country is $349,100km^2$ and includes some parts outside Italian borders and the sea (see Fig. 1). The presence of 124 habitats in the study region was stated in the ISPRA report (ex art. 17 HD), together with the number of quadrats (hereafter denoted M) in which they were present (see Table 1). We used this information as the starting point to construct the sampling strategy for estimating quantities approximating the area of each habitat in the study region.

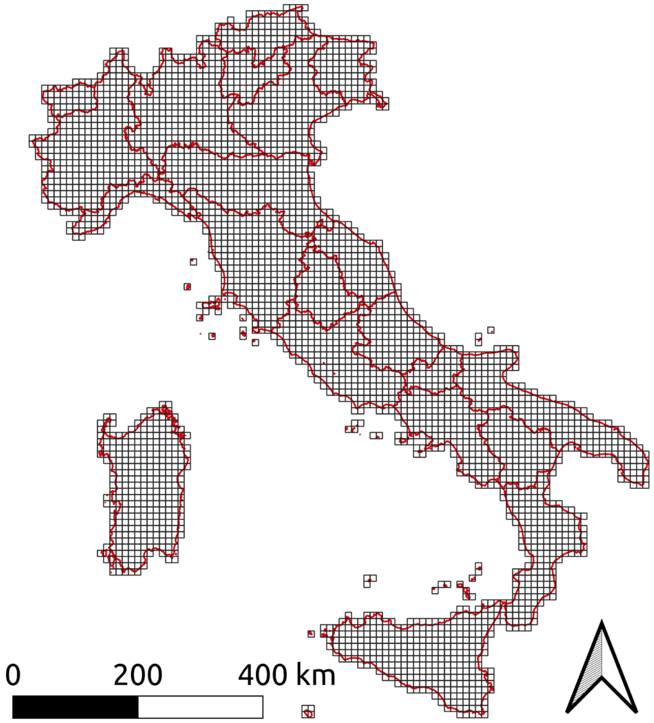

Fig. 1. The study region is covered by $3,491$ quadrats of $10km \times 10km$ (black border and white filling), some of which lie partially outside Italian borders. The administrative borders of the Italian regions are represented by the red line. The map was obtained by intersecting the standard reference grid of quadrats $10km \times 10km$ provided by the European Environmental Agency (European Enviroment Agency, 2013) for the Italian surface area with the administrative borders of the Italian regions (ISTAT, 2022).

Table 1

Number of $10km \times 10km$ quadrats (M) of habitat presence for the 124 habitats present in the study region listed by their HD codes. Habitats highlighted in bold are adopted in the simulation study (see also Fig. 4)

code	*M*	code	*M*	code	*M*	code	*M*
1150	217	3250	392	6410	342	9180	514
1210	552	3260	563	6420	297	9190	18
1240	352	3270	618	6430	937	91AA	1267
1310	222	3280	355	6510	1031	91B0	55
1320	23	3290	221	6520	427	91D0	60
1340	3	4030	303	7110	98	91	1091
1410	247	4060	556	7120	2	91F0	325
1420	199	**4070**	**262**	7140	246	91H0	123
1430	169	4080	236	7150	80	91K0	188
1510	59	4090	140	7210	122	91L0	448
2110	359	**5110**	**48**	7220	307	91M0	619
2120	260	5130	499	7230	318	**9210**	**516**
2130	32	5210	277	7240	62	9220	141
2160	6	5220	16	8110	362	9250	33
2210	193	5230	82	8120	403	9260	1118
2230	322	5310	6	8130	481	**92A0**	**1387**
2240	173	5320	193	8210	1134	92C0	42
2250	212	5330	930	8220	551	92D0	505
2260	133	5410	20	8230	262	9320	247
2270	188	5420	32	8240	160	**9330**	**390**
2330	8	5430	101	8310	757	9340	318
3110	8	6110	436	8320	43	9350	4
3120	74	6130	90	8330	153	9380	41
3130	615	6150	386	8340	128	**9410**	**368**
3140	326	6170	552	9110	402	**9420**	**442**
3150	842	6210	1473	**9120**	**3**	9430	52
3160	39	**6220**	**1566**	9130	334	9510	28
3170	301	6230	548	9140	70	9530	103
3220	394	6240	45	9150	143	9540	195
3230	87	62A0	225	9160	209	9560	6
3240	597	6310	159	9170	1	9580	20

4070 - Bushes with Pinus mugo and Rhododendron hirsutum (*Mugo-Rhododendretum hirsuti*); **5110** - Stable xero-thermophilous formations with Buxus sempervirens on rock slopes (*Berberidion pp*); **6220** - Pseudo-steppe with grasses and annuals of the *Thero-Brachypodietea*; **9120** - Atlantic acidophilous beech forests with Ilex and sometimes Taxus in the shrublayer (*Quercion robori-petraeae or Ilici-Fagenion*); **9210** - Apeninne beech forests with Taxus and Ilex; **92A0** - Salix alba and Populus alba galleries; **9330** - Quercus suber forests; **9410** - Acidophilous Picea forests of the montane to alpine levels (*Vaccinio-Piceetea*); **9420** - Alpine Larix decidua and/or Pinus cembra forests. The entire set of habitat codes along with the related number (M) of $10km \times 10km$ quadrats of habitat presence was extracted from the distribution habitat maps provided by the official Eionet European Central Data Repository (CDR, 2022) for the progress reports and implementation of Article 17(HD).

2.2. Survey arrangement

It was difficult to accurately delineate the ground distribution of most habitats. While recording the size of patches containing the habitat was challenging, recording the presence of the habitat in fixed-area units of adequate size was straightforward. Therefore, we partitioned the M quadrats in which the habitat was present into a grid of $K = 100m \times 100m$ square cells ($1ha$). This cell size was considered a good compromise between the need to perform a complete ecological and cost-effective habitat survey within the cell and an appropriate approximation of the total area. The cells completely outside the Italian borders or completely overlapping with the sea were discarded; the remaining cells constituted the initial population U_0. Moreover, to avoid surveying cells where the habitat presence was impossible, we needed to quantify the chance of habitat presence in the cells. Therefore, for each cell $j \in U_0$, the ISPRA Group for Terrestrial Habitat Monitoring and Conservation calculated a value $x_j \geqslant 0$, referred to as the habitat suitability score (HSS). Scores equal to 0 were assigned to cells where the habitat presence was impossible, and these cells were discarded (see Section 2.3). Then, the target population to be surveyed was constituted by the set $U \subset U_0$ of N cells in which $x_j > 0$, i.e., habitat presence was considered possible (see Fig. 2).

The target parameter under estimation was established to be the number of cells Y in which the habitat is present. In practice, Y constitutes the total area of the cells that cover the habitat at a grain size of $1ha$, and as such, it is referred to as habitat coverage. To estimate Y, we introduced a survey variable indexing the presence/absence of the habitat in the cells, i.e., for each cell $j \in U, y_j$ was set to be 1 if the habitat was present in the cell and 0 otherwise. In this way, the target quantity Y was the population total, i.e.,

$$Y = \sum_{j \in U} y_j.$$

We aimed to select samples with cells evenly spread throughout the region of habitat presence, thus achieving spatial balance (e.g., Grafström and Lundström, 2013; Brown et al., 2015). Moreover, to maximize the likelihood of encountering the habitat within the selected cells, HSS values were adopted as auxiliary information to guide cell sampling. Owing to the large effort that may be required for detecting the habitat presence with $1ha$ cells, we established a maximum sampling fraction of 0.04%.

A further source of auxiliary information was the habitat presence in some cells from recent investigations. In particular, we used recently available habitat maps representing spatial polygons with previously validated habitat presence (see the methodology for producing the habitat maps in Carli et al., 2020) and the spatial polygons representing all the Italian protected sites within the N2K network (MiTE, 2021). The N2K is the largest coordinated network of protected areas in the world, extending across all the 28 EU countries and designated under the HD. Specifically, for each cell $j \in U, h_j$ was set to 1 if it was known that the habitat was present in the cell and 0 otherwise. Accordingly, the population total of this variable was

$$H = \sum_{j \in U} h_j$$

and constituted the extent of the region in which habitat presence was certain and was the lower bound for any estimate of Y.

2.3. Determination of habitat suitability scores (HSSs)

HSSs were determined based on a set of variables correlated with the presence/absence of the habitat in the cells (i.e., survey variable) and available for all cells in the study region. Thus, from variables for the habitat characterization provided by the official manual for habitat monitoring (Angelini et al., 2016) and available (informatic layers) at the national scale, ISPRA experts first selected the following (Table 2): (i) land use types (CLC, 2018), (ii) exposure (derived from DEM20), (iii) altitude (derived from DEM20 - SINAnet 2020), (iv) slope (derived from DEM20), (v) hydrographic network (HydRet - SINAnet 2020) and (vi) distance to the coastline (ISTAT, 2020). Each predictor included a set of classes derived from the structure of the data (e.g., CLC, 2018) or established based on expert knowledge at the national scale (e.g., distance to the coastline for coastal habitats). For each habitat type, experts assigned one of the following weights to each predictor class: “0” (null suitability for habitat presence), “0.5” (intermediate suitability for habitat presence), or “1” (high suitability for habitat presence).

For each cell $j \in U_0$, we then obtained an HSS score $0 \leqslant x_j \leqslant 1$ multiplying all the weights associated with each predictor such that a score resulted in 0, i.e., no possibility of habitat presence, if at least one of the weights was 0. All the predictors listed in Table 2 were used to ecologically characterize the habitat types adopted for this study. Notably, if the predictor was considered ecologically irrelevant to define the distribution of a specific habitat type, we assigned a value of 1 to all categories of the ”irrelevant variable”, thus not affecting the final suitability score (as the suitability score is the result of multiplication, and multiplying by 1 does not change the suitability score). To improve

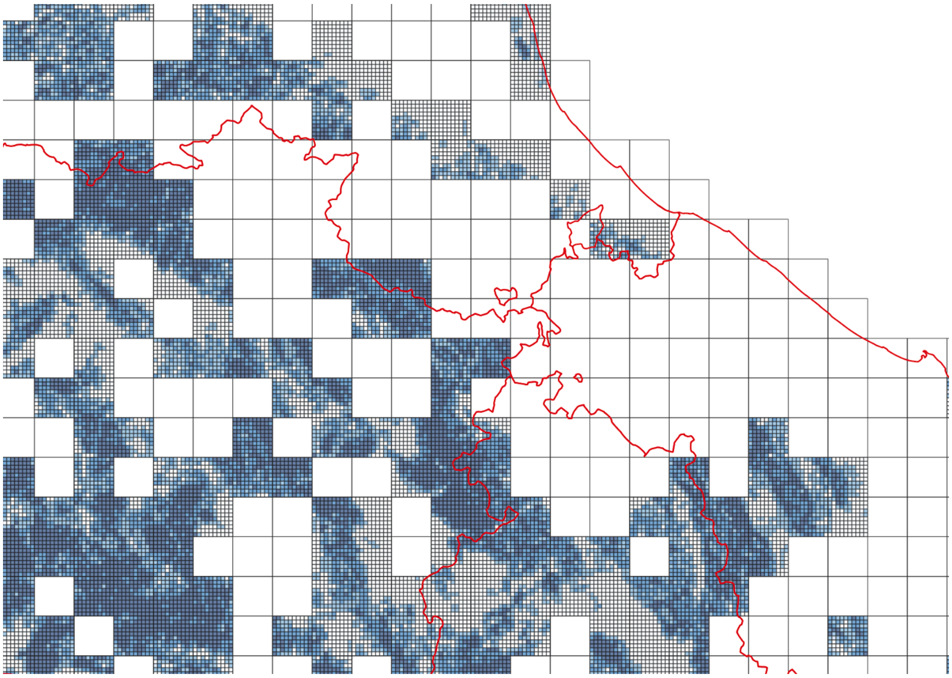

Fig. 2. An example of the population of cells with HSSs score > 0 to be sampled. The colour scale from light blue to blue indicates the increasing value of the HSS scores. White-coloured cells had an HSS equal to 0 and were discarded. The administrative borders of the Italian regions are represented by the red line.

Table 2
Summary of the environmental predictors selected for determining habitat suitability scores (HSSs). The reported on-line data sources were visited in July 2020.

Description	Type	Data source
Corine Land Cover	Raster	CLC 2018. Version is v.2020_20u1. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download
Digital Elevation Model	Raster	DEM20. Digital Elevation Model. Rete del Sistema Informativo Nazionale Ambientale. SINAnet. http://www.sinanet.isprambiente.it/it/sia-ispra/download-mais/dem20/view
Hydrographic network	Shape file	HydRet. Hydrographic network. Rete del Sistema Informativo Nazionale Ambientale. SINAnet. http://www.sinanet.isprambiente.it/it/sia-ispra/download-mais/reticolo-idrografico/view
Coastal Line	Shape file	ISTAT 2020-Sezioni di Censimento Litoranee. Linea litoranea. https://www.istat.it/it/archivio/137341

the chance of sampling cells in which the habitat was present we increased the HSS of cells falling within a known habitat coverage polygon (Carli et al., 2020) by multiplying it by 8. Moreover, to further improve the chance of sampling cells within the N2K network polygons, their HSS value was multiplied by 1.25. Consequently, the HSS of a cell located both within a habitat coverage polygon and a N2K polygon that varied originally from 0 to 1 was multiplied by 8 and then by 1.25. After these rescaling operations, HSSs were in the range 0 −10.

2.4. Sampling and estimation

Cell sampling was performed separately for each habitat. We adopted a two-phase scheme, with the first phase comprising two stages (Sections 1–4 of the Supporting Information file). The complexity of the scheme was due to the necessity of a final sample of cells evenly distributed among and within the quadrats and having high HSS values.

The first stage of the first phase was performed to evenly distribute selected quadrats throughout the study region. For this purpose, the set of the M quadrats in which the habitat was present was partitioned into m clusters of neighbouring quadrats, referred to as the q-blocks. The number of quadrats per q-block was established to ensure that blocks had approximately the same number of quadrats. Partition was performed adopting the k-means algorithm (k-means clustering - “*stats*” package R Core Team, 2020). This algorithm was originally proposed by Hartigan and Wong, 1979. For this purpose, the algorithm needed the number of clusters m, the spatial coordinates of the quadrat centroids and the maximum number of iterations allowed, which was established to be $100,000$. Subsequently, in accordance with the sampling scheme referred to as the one-per-stratum sampling (e.g., Breidt, 1995), henceforth OPSS, a quadrat was selected in each block with probabilities proportional to the HSS totals within blocks. Because a unique quadrat was selected within clusters of contiguous quadrats, OPSS ensured that samples of quadrats were evenly spread throughout the study region. Moreover, because the selection was performed with probabilities increasing with the HSS totals within quadrats, the quadrats with high HSSs had a greater chance of being selected.

The number m of quadrats to select from the M was established by the following function of M

$$m = \begin{cases} M & if \quad M \leqslant 10 \\ [9.0756303 + 0.0924369M] + 1 & if \quad 10 < M < 1200 \\ [0.1M] & if \quad M \geqslant 1200 \end{cases} \tag{1}$$

where $[x]$ is the integer part of x. The algorithm was used to adjust the sampling effort with respect to the total number of quadrats M, avoiding excessive effort when the number of quadrats was large. In particular, the algorithm established that no selection was performed if the number of quadrats was smaller than 10, in which case all quadrats were included in the sample, while the percentage of selected quadrats decreased linearly from 100% when $M = 10$ to 10% when $M = 1200$, remaining equal to 10% for any M greater than 1200.

The second stage of the first phase was performed to evenly spread the selected cells within the quadrats selected in the first stage. For this purpose, OPSS was once again performed within the selected quadrats. Because the cells were arranged in a regular grid of size $100m \times 100m$, there was no need for time consuming clustering algorithms to determine clusters of neighbouring cells. The m quadrats selected in the first stage were partitioned into $k = 25$ quadrat blocks of 20×20 cells, referred to as c-blocks (see Fig. 3). The blocks where the habitat was absent were discarded, and a cell was randomly selected within each of the remaining c-blocks in accordance with the OPSS scheme with probabilities proportional to the HSSs to ensure that cells with high HSSs had a higher chance of being selected. Therefore, a maximum of 25 cells was selected within each selected quadrat at the end of the first phase.

Finally, the second phase was performed to reduce the sampling effort from a maximum of $k = 25$ cells per quadrat to a maximum of $\overline{n} = 4$ cells. Because the spatial balance of selected cells within quadrats was already achieved by the use of OPSS in the second stage of the first phase, the spatial component was ignored and only the HSSs were considered to ensure that cells with high HSSs had a higher chance of being selected. Accordingly, a sample of $\overline{n} = 4$ cells was selected with probability proportional to the HSS of the cells selected in the first phase within each quadrat. Selection was performed with the Sampford algorithm (Sampford, 1967). If the cells selected in a quadrat at the end of the first phase were less than or equal to 4, second-phase selection was not carried out, and all the cells were included in the final sample.

Once the final sample was achieved, Y was estimated by means of the double expansion (DE) estimator (e.g., Särndal et al., 1992, Section 9.3) $\widehat{Y}_{(2)DE}$ given by equation (SM.9). The DE estimator was adopted because it is able to handle the complexities involved in using multi-phase sampling schemes. The DE estimator was design-unbiased with the design-based variance given by equation (SM.11). If the habitat presence was known for some cells in the population, we exploited this additional information by means of the difference (DIF) estimator (e.g., Särndal et al., 1992, Section 6.3) $\widehat{Y}_{(2)DIF}$ given by equation (SM.13). The DIF estimator was simply a modification of the DE estimator to include the additional information in the estimation criterion. It has been recently used in biodiversity surveys to improve species richness estimation (Chiarucci et al., 2018). The DIF estimator was design-unbiased with design-based variance given by equation SM.14. In our case, the DIF estimator had the appealing property to providing consistent results in that the estimate was never smaller than the number of cells H in which the habitat presence was known, which obviously should constitute a lower bound for any estimator. This feature was not ensured by the DE estimator.

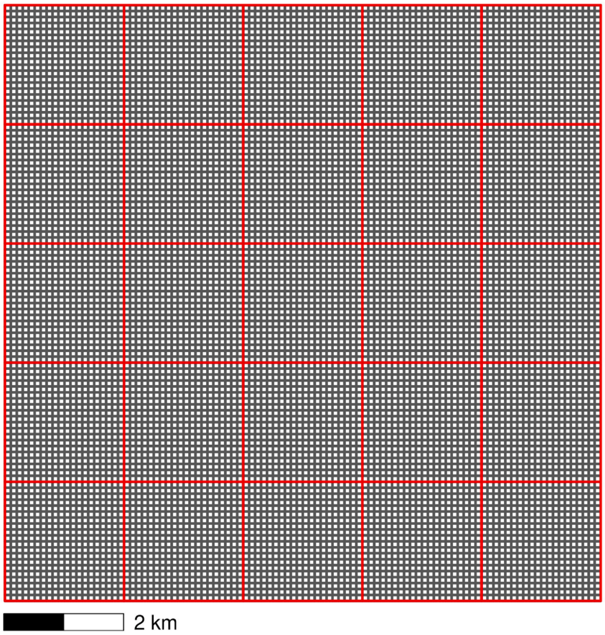

Fig. 3. First stage selected quadrat partitioned into $k = 25$ square c-blocks (c-block edges in red) of 400 cells (cell edges in black).

Notably, auxiliary information from which to construct HSSs and previous knowledge of the cells with habitat presence are not essential for executing the strategy. If no auxiliary information is available, all the HSSs are set to be equal such that all the quadrats have an equal chance to be selected, and only the even spread of the selected quadrats and cells is ensured by the OPSS. Moreover, if no previous knowledge of habitat presence is available, all the h_j are set equal to 0 such that DIF and DE estimators coincide.

The variances of the DE estimator $\widehat{Y}_{(2)DE}$ and the DIF estimator $\widehat{Y}_{(2)DIF}$, were estimated using the Hansen–Hurvitz (HH)-like variance estimators (e.g., Wolter, 2007) V_{DE}^2 and V_{DIF}^2 according to equations (SM.12) and (SM.15), respectively.

2.5. Nonresponse treatment

For cells that are impossible to reach in the field, due to topographic constraints, it was impossible to verify the habitat presence. We thus treated these missing values as nonresponses, and we followed the design-based suggestion by Fattorini et al., 2013, i.e., we corrected the DE and DIF estimators performed on the respondent sample by nonresponse calibration weighting (Haziza et al., 2010). The purpose was to increase the estimates achieved from the respondent sample to reduce the downwards bias invariably induced by nonresponses. By this approach, nonresponses were viewed as fixed characteristics of the cells, without attempting any model to explain them. This approach was performed by introducing for each cell a response indicator z_j that was equal to 1 if it was possible to reach and explore and 0 otherwise (see Section 7 in the Supporting Information file for information on nonresponse treatment in sample surveys).

If the DE criterion was adopted, we then calibrated the estimator computed on the respondent sample by the estimator $\widehat{Y}_{(2)DE-CAL}$ according to equation (SM.19). If the DIF criterion was adopted, calibration was performed by the estimator $\widehat{Y}_{(2)DIF-CAL}$ in equation (SM.23). Section 7 in the Supporting Information file provided the conditions under which the two calibrated estimators reduced the downwards bias and turned out to be approximately unbiased. The condition was that the relationships between the y_js and x_js in the case of the estimator $\widehat{Y}_{(2)DE-CAL}$ or that between the d_js and x_js in the case of $\widehat{Y}_{(2)DIF-CAL}$ were similar in respondent and nonrespondent cells (see also Fattorini et al., 2013).

The variances of $\widehat{Y}_{(2)DE-CAL}$ and of $\widehat{Y}_{(2)DIF-CAL}$ were estimated by the HH-like variance estimators V_{DE-CAL}^2 and $V_{DIF-CAL}^2$ according to equations (SM.22) and (SM.24), respectively.

2.6. Coverage estimation within subregions

Coverage estimation was performed for the entire Italian surface area and for the three biogeographical regions partitioning the study region (Cervellini et al., 2020) as well as for the portion of cells located within the Natura 2000 Network. For this purpose, we introduced for each cell $j \in U$ an indicator u_{gj} that was equal to 1 if the cell was in the subregion g of interest and 0 otherwise. Then, we adopted the estimator $\widehat{Y}_{(2)DE}$ or $\widehat{Y}_{(2)DIF}$ in the case of complete samples or the estimator $\widehat{Y}_{(2)DE-CAL}$ or $\widehat{Y}_{(2)DIF-CAL}$ in the case of nonresponses together with their corresponding variance estimators simply by multiplying the y_js by u_{gj}s. This strategy has been widely adopted in sample surveys when estimating totals in particular sectors of populations and is usually referred to as domain estimation (see Särndal et al., 1992, Chapter 10 and Section 8 in the Supplementary Information file for more details).

2.7. Simulation study

A simulation study was performed to check and compare the precision of the estimation strategies corresponding to the estimators $\widehat{Y}_{(2)DE}$ and $\widehat{Y}_{(2)DIF}$. To be realistic, we selected nine habitats from those listed in Table 1 for which HSS values were available. The selected habitats showed different spatial distributions, representing the three levels of spatial occurrences within quadrats (i.e., $M \leq 10, 10 < M < 1200$ and $M \geq 1200$, see Algorithm 1), with the very rare habitat 9120, the scattered habitat 5110, the more common habitats 4070, 9210, 9330, 9410 and 9420 and the very common habitats 6220 and 92A0 (see Fig. 4).

For these habitats, we standardized HSSs in the range $0-1$ by dividing each HSS by 10. The number of cells N with HSS values greater than 0 indicated the territory (in *ha*) where the habitat presence was possible, i.e., the extent of the survey area. Based on these territories and from previous information on habitat validated presence (Carli et al., 2020), we attempted to provide realistic coverages for each habitat, establishing the number of cells Y where the habitat was present. The resulting coverages ranged from a minimum of 0.4% of the survey area for habitat 9120 to a maximum of 7.9% for habitat 9410 (see Table 3).

We then generated habitat presence by sorting the cells with respect to their HSSs from the greatest to the smallest value and assigning $y_j = 1$ to the first Y cells and $y_j = 0$ to the remaining $N - Y$. In practice, we generated presence in cells with the greatest HSSs. Moreover, knowledge of habitat presence from past investigations was incorporated by assigning $h_j = 1$ to cells for which $y_j = 1$, which were also present in the ISPRA polygons, and assigning $h_j = 0$ to the remaining cells. Once the y_js and h_js were generated, their totals Y and H were determined. All these values remained fixed throughout the simulation runs as fixed characteristics because in design-based approaches, uncertainty stems only from sampling.

For each habitat, we independently performed $R = 100,000$ two-phase selections of cells following the sampling scheme described in subSection 2.4. In the second phase, we selected a maximum number of $\overline{n} = 1, 2, 3, 4$ cells. Because the final sample size n was a random variable depending on the number of cells with positive HSSs within the selected c-blocks, the expected sample size (ESS) was empirically computed as

$$ESS = \frac{1}{R}\sum_{r=1}^{R} n_r$$

where n_r is the size of the final sample selected at the r-th simulation run. Moreover, the expected fraction of habitat presence in the sample (EPS) was empirically computed as

$$EPS = \frac{1}{R}\sum_{r=1}^{R} \frac{H_r}{n_r}$$

where H_r is the number of cells with habitat presence in the final sample. The EPS values were compared with those expected under simple random sampling without replacements (SRSWOR) that coincided with the fraction of cells with habitat presence in the population, i.e., $p = Y/N$.

Moreover, for each sample, the estimators $\widehat{Y}_{(2)DE}$ and $\widehat{Y}_{(2)DIF}$ were computed from the sample data together with their variance estimators V^2_{DE} and V^2_{DIF}. At the end of the procedure, for each habitat and both estimators, we determined R coverage estimates, $\widehat{Y}_1, \ldots, \widehat{Y}_R$, and the corresponding variance estimates, i.e., $V^2_1, \ldots, V^2_R$, from which we derived the relative standard error estimates $R\widehat{S}E_1, \ldots, R\widehat{S}E_R$ with $R\widehat{S}E_r = V_r/\widehat{Y}_r$ for $r = 1, \ldots, R$. Finally, the confidence interval at the nominal level of 0.95 was achieved by $\widehat{Y}_r \pm 2V_r$. For each habitat and each estimator, the two collections constituted the Monte Carlo distributions of the abundance estimator and of its relative standard error estimators. The collections mimicked the unknown corresponding distributions and were adopted to empirically determine the theoretical properties.

Accordingly, from the resulting Monte Carlo distributions achieved by simulation, the expectation and the variances of the abundance estimator were empirically determined as follows:

$$E = \frac{1}{R}\sum_{r=1}^{R} \widehat{Y}_r$$

and

$$Var = \frac{1}{R}\sum_{r=1}^{R} V^2_r.$$

From these quantities, the relative bias $RB = (E - Y)/Y$ and the relative standard error $RSE = \sqrt{Var}/Y$ were determined. We then tested the design effect (e.g., Särndal et al., 1992, Section 2.10) in terms of the RSE. In practice, the RSEs of the two estimators were compared with those achieved by the Horvitz-Thompson (HT) estimator under SRSWOR with sample size ESS, i.e.,

$$RSE_{SRSWOR} = \sqrt{\frac{N - ESS}{N \times ESS}\frac{1-p}{p}}.$$

Moreover, the expectation of the relative standard error estimators was achieved as follows:

$$ERSEE = \frac{1}{R}\sum_{r=1}^{R} R\widehat{S}E_r.$$

Finally, the actual coverage of the nominal 0.95 confidence intervals $C95$ was obtained as the fraction of the intervals containing the true coverage Y. Additionally, both estimators ensured design-unbiasedness. Therefore, their RB values were theoretically known to be 0, and their empirical counterpart RBs were considered only to confirm the reliability of the simulation study.

To check the effectiveness of bias reduction in the presence of nonresponses, a further simulation study was performed. Nonresponses were artificially generated from the populations adopted in the previous study assigning a dichotomous index $r_j = 1$ if it was possible to reach the cell j and $r_j = 0$ otherwise. Nonresponses were considered impossible, i.e., $r_j = 1$ for those cells such that $h_j = 1$, i.e., cells where habitat presence was known from previous investigations, which obviously implied

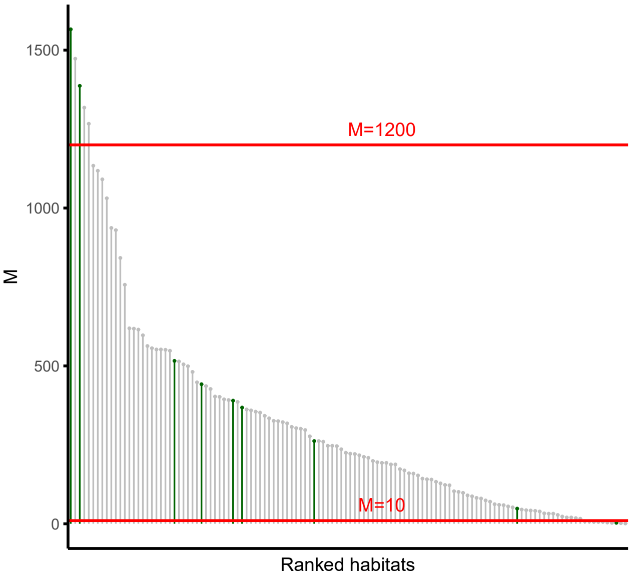

Fig. 4. Bar graph showing the ranked M for the 124 habitats in the study region reported in Table 1. Horizontal lines denote the three levels of occurrence ($M \leq 10, 10 < M < 1200$ and $M \geq 1200$). Habitats highlighted in green were adopted in the simulation study.

Table 3

Monte-Carlo performance of double expansion and difference estimators of coverage compared with the Horvitz-Thompson estimator under simple random sampling and performance of relative standard error estimators for nine habitats in the study region

Habitat	$\overline{n}$	ESS	EPS (%)	RSE (%)			ERSEE (%)		C95 (%)	
				SRSWOR	DE	DIF	DE	DIF	DE	DIF
4070 survey area=1,948,964*ha*	1	34.0	33.5	70.1	31.4	28.8	38.7	21.1	94.0	73.5
coverage=110,000 ha (5.6%)	2	68.0	33.5	49.6	23.4	20.6	28.3	17.9	94.9	81.4
known coverage=72,559*ha* (66.0%)	3	102.0	33.0	40.5	20.2	17.1	23.5	15.6	95.1	85.0
ss mean=0.10	4	136.0	32.3	35.1	18.4	15.0	20.5	13.9	94.9	86.8
corr. presence/ss=0.88										
$M = 262$, $m = 34$										
5110 survey area=446,098*ha*	1	14.0	7.1	281.0	165.6	163.8	66.2	32.3	65.3	43.5
coverage=4,000*ha* (0.9%)	2	28.0	7.1	198.7	119.5	117.0	71.3	38.8	75.2	60.7
known coverage=1,225*ha* (30.6%)	3	42.0	7.1	162.2	95.8	92.9	67.4	40.3	77.9	66.9
ss mean=0.05	4	56.0	6.9	140.5	83.7	80.6	62.3	39.9	82.7	70.3
corr. presence/ss=0.84										
$M = 48$, $m = 14$										
6220 survey area=12,820,526*ha*	1	157.0	33.7	47.6	14.5	14.0	17.5	15.6	97.0	95.7
coverage=350,000*ha* (2.7%)	2	313.9	33.6	33.7	10.8	10.2	12.4	11.1	97.1	96.1
known coverage=87,465*ha* (25.0%)	3	470.4	33.2	27.5	9.3	8.6	10.2	9.1	96.5	95.9
ss mean=0.03	4	626.6	32.3	23.9	8.4	7.7	8.8	7.9	96.0	95.5
corr. presence/ss=0.89										
$M = 1566$, $m = 157$										
9120 survey area=28,993*ha*	1	3.0	17.7	877.4	122.6	43.0	43.1	0.4	46.3	0.5
coverage=125*ha* (0.4%)	2	6.0	17.6	620.4	85.9	32.3	63.1	0.9	73.6	1.2
known coverage=121*ha* (98.8%)	3	9.0	17.6	506.5	67.2	24.8	69.3	1.0	89.1	1.6
ss mean=0.08	4	12.0	15.2	438.6	59.5	21.8	66.7	1.3	90.6	2.3
corr. presence/ss=0.84										
$M = 3$, $m = 3$										
9210 survey area=5,058,912*ha*	1	57.0	31.1	59.4	20.8	13.7	28.7	5.9	97.9	42.5
coverage=240,000*ha* (4.7%)	2	114.0	31.2	42.0	16.1	9.8	20.5	5.8	97.8	57.7
known coverage=221,279*ha* (92.2%)	3	171.0	31.1	34.3	14.1	8.0	16.8	5.4	97.1	65.0
ss mean=0.12	4	228.0	30.5	29.7	13.1	7.0	14.6	5.1	96.4	69.2
corr. presence/ss=0.97										
$M = 516$, $m = 57$										
92A0 survey area=12,683,736*ha*	1	139.0	26.2	83.3	17.3	13.7	18.7	13.2	95.5	91.1
coverage=130,000*ha* (1.0%)	2	278.0	26.3	58.9	13.3	10.0	13.3	9.6	94.7	92.6
known coverage=77,418*ha* (59.6%)	3	417.0	25.9	48.1	11.7	8.5	10.9	7.9	93.2	92.6
ss mean=0.07	4	556.0	24.9	41.7	10.7	7.5	9.5	6.9	91.8	92.4
corr. presence/ss=0.80										
$M = 1387$, $m = 139$										
9330 survey area=3,414,940*ha*	1	46.0	14.9	127.6	46.0	15.7	47.7	10.0	87.6	67.3
coverage=45,000*ha* (1.3%)	2	92.0	14.9	90.2	36.7	12.1	34.3	8.1	88.2	74.7
known coverage=38,930*ha* (86.5%)	3	138.0	14.7	73.7	32.9	10.8	28.2	7.1	87.2	76.8
ss mean=0.11	4	184.0	14.2	63.8	31.0	10.0	24.6	6.5	85.4	78.0
corr. presence/ss=0.57										
$M = 390$, $m = 46$										
9410 survey area=3,312,339*ha*	1	44.0	49.3	51.7	18.8	16.3	26.7	11.7	98.2	74.0
coverage=260,000*ha* (7.9%)	2	88.0	49.4	36.5	13.9	11.6	19.1	9.6	98.6	81.7
known coverage=210,884*ha* (81.1%)	3	132.0	49.0	29.8	11.8	9.5	15.8	8.3	98.4	84.7
ss mean=0.14	4	176.0	48.1	25.8	10.6	8.3	13.7	7.5	98.3	86.7
corr. presence/ss=0.95										
$M = 368$, $m = 44$										
9420 survey area=3,874,989*ha*	1	51.0	39.7	69.8	21.3	17.1	25.7	13.3	96.4	77.1
coverage=150,000*ha* (3.9%)	2	102.0	39.7	49.3	16.1	12.3	18.5	10.8	96.5	83.9
known coverage=117,649*ha* (78.4%)	3	153.0	39.3	40.3	13.8	10.1	15.2	9.2	96.1	86.8
ss mean=0.10	4	204.0	38.4	34.9	12.6	8.9	13.3	8.2	95.5	
corr. presence/ss=0.91										
$M = 442$, $m = 51$										

ESS=expected sample size, EPS=expected presence in the sample (%), RSE=relative standard error (%), ERSEE=expectation of relative standard error estimator (%), C95=coverage of the 0.95 confidence interval (%), DE=double expansion estimator and DIF=difference estimator. Purple values refer to RSEs (%) achieved by the HT estimator under simple random sampling without replacement (SRSWOR) with ESS taken to have a fixed sample size. Values in blue and green refer to DE and DIF estimation, respectively.

the possibility of reaching them. For the remaining cells with $h_j = 0$, nonresponses were established by generating a random number u uniformly distributed in the interval $(0, 1)$ and then assigning $r_j = 0$ if u was smaller than a previously established nonresponse rate ρ and $r_j = 1$ otherwise. To consider several levels of nonresponses, simulations were performed for $\rho = 0.05, 0.10, 0.20$. Once the r_js were generated, they remained fixed throughout the simulation runs as fixed characteristics of the cells because of the design-based nature of the nonresponse treatment adopted in this study (see Section 7 of the Supporting Information file).

For each habitat and each nonresponse level ρ, we independently performed $R = 100,000$ two-phase selections of cells following the sampling scheme described in subSection 2.4. In the second phase, we selected a maximum number of $\overline{n} = 4$ cells as the most sustainable effort. Because the number of nonrespondent cells in the final sample was a random variable, the expected number of nonresponses (ENR) was empirically computed as

$$ENR = \frac{1}{R}\sum_{r=1}^{R} nor_r$$

where nor_r is the number of nonresponses in the final sample selected at the r-th simulation run.

Moreover, for each selected sample, the estimators $\widehat{Y}_{(2)DE-CAL}$ and $\widehat{Y}_{(2)DIF-CAL}$ were computed from the sample data together with their variance estimators V^2_{DE-CAL} and $V^2_{DIF-CAL}$. For each habitat, each nonresponse rate and both estimators, we achieved the collection of the R coverage estimates, $\widehat{Y}_1, \ldots, \widehat{Y}_R$, and the corresponding variance estimates, $V^2_1, \ldots, V^2_R$, from which we derived the relative standard error estimates $R\widehat{S}E_1, \ldots, R\widehat{S}E_R$ with $R\widehat{S}E_r = V_r / \widehat{Y}_r$ for $r = 1, \ldots, R$. Finally, the confidence interval at the nominal level of 0.95 was achieved by $\widehat{Y}_r \pm 2V_r$. Then, from the resulting Monte Carlo distributions, we derived the expectation and the mean squared errors of the abundance estimator as follows:

$$E = \frac{1}{R}\sum_{r=1}^{R} \widehat{Y}_r$$

and

$$MSE = \frac{1}{R}\sum_{r=1}^{R} \left(\widehat{Y}_r - Y\right)^2.$$

From these quantities, the relative bias $RB = (E - Y)/Y$ and the relative root mean squared error $RRMSE = \sqrt{MSE}/Y$ were determined together with the expectation of the relative standard error estimators

$$ERSEE = \frac{1}{R}\sum_{r=1}^{R} R\widehat{S}E_r$$

and the actual coverage of the nominal 0.95 confidence intervals $C95$ as the fraction of intervals containing the true coverage Y. Notably, neither estimator ensured design-unbiasedness. Therefore, their RB values were most important and their precision was quantified by their MSEs rather than by their variances.

3. Results

The values of relative bias (RB) concerning the double expansion (DE) and difference (DIF) estimators—theoretically equal to 0—were very close to 0 (always smaller than 0.7%).

The generated populations showed standardized HSS mean values of approximately 0.10 for most habitats except for habitat 6220 which showed approximately 0.03, and habitat 9410 which showed approximately 0.14. The way in which habitat presence was generated within cells gave rise to a strong correlation between x_js and y_js that varied between 0.80 and 0.97, except for habitat 9330, which showed a correlation of 0.57. Of the cells with habitat presence, the percentage of cells where habitat presence was known varied from 25% for habitat 6220 to 99% for habitat 9120 (Table 3).

The increase from 1 to 4 of the maximum number of cells to select in the second phase within the selected quadrats ($\overline{n}$) led to considerable increases in the precision with RSEs of DE that halved their values in most cases. However, the reduction in RSEs decreases as $\overline{n}$ increases, suggesting that further increments of $\overline{n}$ over 4 are unsuitable, increasing the sampling effort without producing relevant improvements (Table 3 and Fig. 5).

The expected sample sizes (ESS) were invariably equal to $\overline{n}m$, showing that it was highly improbable to find quadrats where the number of c-blocks available for the second-phase sampling was smaller than $\overline{n}$. Of the total number of selected cells, the expected percentages of selected cells with habitat presence were much greater—from approximately 6 to approximately 40 times—than those expected under SRSWOR (Table 3).

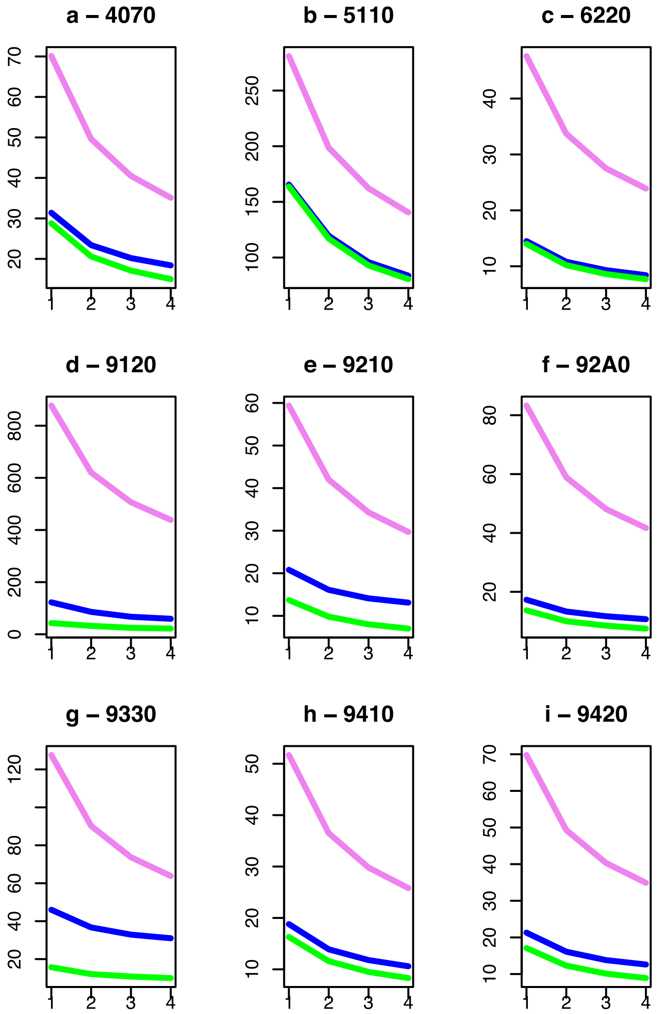

Fig. 5. Graphs of the relative standard error (RSE) in percentage plotted against $\overline{n}$ from 1 to 4 for each habitat and for the Horvitz-Thompson estimator (HT) under simple random sampling without replacement (SRSWOR) and for the DE and DIF estimators under the proposed scheme (purple, blue and green lines, respectively, for the values in Table 3).

The precision of the DE estimator with respect to the crude HT estimator under SRSWOR was high, with efficiencies (ratio of RSEs) that varied from approximately 1.5 in habitat 5110 to approximately 7 in habitat 9120. In most cases, efficiencies ranged from 2 to 3 (Table 3 and Fig. 5).

The DIF estimator that exploited the previous knowledge of habitat presence in unsampled cells invariably outperformed the DE estimator (Table 3 and Fig. 5). Gains in precision were relevant when previous knowledge of habitat presence covered a large percentage, over 60%, of the habitat coverage (Fig. 5 a, d, e, g, h, i).

In absolute terms, for $\overline{n} = 4$, i.e., sampling fractions always smaller than 0.04%, the proposed sampling scheme combined with the DIF estimator yielded suitable precision with RSEs smaller than 15% when (i) populations were large (some millions of cells), (ii) HSSs were good proxies for habitat presence with correlations of x_js vs y_js of approximately 0.8–0.9, and (iii) the habitat coverage with respect to the survey areas was not smaller than 1%. These features were shared by habitats 4070, 6220, 9210, 92A0, 9410 and 9420. In these cases, RSEs were approximately 7–8%, except for habitat 4070, for which the RSE was 15% (Fig. 5 a, c, e, f, h, i). For habitats 5110, 9120 and 9330, the precision was unsuitable (Fig. 5 b, d, g). Particularly for habitat 5110, the RSE of the DE estimator was 83%, and the DIF estimator did not yield relevant improvements (Fig. 5 b).

The tentative RSE estimator that applied the HH criterion to the cells selected in the second phase using the product of the first-phase inclusion probabilities with those of the second phase as if they were the actual first-order inclusion probabilities provided unsatisfactory results, especially for the DIF estimator. While the proposed estimator performed quite well for the DE estimator, in most cases providing moderate overestimation with coverages of confidence intervals near the nominal level of 95%, it invariably underestimated the RSEs of the DIF estimator, with interval coverages invariably smaller than the nominal level. In some cases (e.g., habitat 9120), the underestimation was unsuitably large (Table 3). Therefore, as a precautionary rule of thumb, the RSE estimates achieved for the DE estimator should also be used to estimate the RSEs of the DIF estimator.

Since the RSEs of the DIF estimator were always smaller than those of the DE estimator, which in turn were overestimated, a fortiori these estimates should also overestimate the RSEs of the DIF estimator.

For the simulation study performed to check the nonresponse effects on the properties of the estimator, estimation based on the respondent sample was equivalent to estimation based on the complete sample when the sample values that could not be recorded were set to 0. Therefore, the effect of nonresponses turned out to be negligible if most of them occurred where the habitat was absent. Moreover, given that nonresponses were not allowed where the habitat presence was known from previous investigations, nonresponse effects were weak for habitats with high percentages of known coverages such as 9120, 9210 and 9330, where these percentages were greater than 85%. In these cases, even under a nonresponse rate of 20%, the negative bias was smaller than 3%, and the DIF estimator performed on the respondent samples provided the best results in terms of RRMSEs (Table 4).

In the other cases, when nonresponses also occurred where the habitat was present, they heavily impacted the bias of both the DE and DIF estimators, with biases in some cases reaching levels of −15%. The bias that affected the estimators in the case of nonresponses decreased the precision, producing RRMSEs that were greater than the RSEs achieved with complete samples (Tables 3 and 4). At the same time, bias increased the underestimation of RRMEs below the actual values and skewed the confidence intervals, decreasing their actual coverages. In these cases, calibration was necessary. However, the DE-CAL estimator eliminated negative bias at the cost of inducing a positive bias that sometimes was greater (in absolute value) than that entailed by nonresponses. On the other hand, the DIF-CAL estimator considerably reduced the negative bias even without reversing the sign and produced RRMSEs invariably smaller than those produced by DE-CAL. Therefore, the use of DIF-CAL seemed to be the best solution when calibration is necessary.

However, as in the case of complete samples, the HH-like variance estimator (SM.24) underestimated the actual precision, producing poor coverage of the resulting confidence intervals. In such cases, as a precautionary rule of thumb, the estimator (SM.22) achieved for the DE-CAL estimator should also be used to estimate the precision of the DIF-CAL estimator and to construct confidence intervals.

4. Discussion

We developed an adaptive (sensu Lindenmayer and Likens, 2009, 2018; Lindenmayer et al., 2020) sampling strategy to provide unbiased estimators, or nearly unbiased in the presence of nonresponses, of habitat coverage over the study region (expressed as the number of 1*ha* cells occupied by habitat type) as an affordable and sound measure to satisfy the quantitative measurement of the "area" criterion in

Table 4

Monte-Carlo performance of double expansion and difference estimators of coverage calibrated to account for nonresponses and performance of relative standard error estimators for nine habitats in the study region and three nonresponse rates compared with the same indicators achieved from the respondent sample neglecting nonresponse presences

			RESPONDENT SAMPLE								CALIBRATION							
			RB (%)		RRMSE (%)		ERSEE (%)		C95 (%)		RB (%)		RRMSE (%)		ERSEE (%)		C95 (%)	
Habitat	NRR (%)	ENR	DE	DIF	DE	DIF	DE	DIF	DE	DIF	DE	DIF	DE	DIF	DE	DIF	DE	DIF
4070	5	5.1	−1.7	−1.7	18.3	19.9	20.6	17.9	94.0	91.2	2.4	−0.5	19.9	15.2	17.9	13.1	91.2	85.2
	10	10.2	−3.4	−3.4	18.4	14.8	20.7	15.5	92.9	81.9	4.4	−1.0	20.5	15.4	17.2	12.8	89.7	83.3
	20	20.2	−6.9	−6.9	19.0	15.2	20.9	13.1	90.3	75.4	8.8	−2.4	22.5	15.8	15.8	12.0	84.8	78.8
5110	5	2.8	−3.9	−3.9	81.8	78.5	62.8	59.0	81.3	68.1	1.4	−0.4	84.4	80.6	59.0	37.7	81.8	69.0
	10	5.5	−7.2	−7.2	80.5	77.2	63.4	38.5	79.8	65.8	3.1	−0.4	87.1	83.0	56.4	35.7	80.8	67.3
	20	11.0	−14.1	−14.2	78.0	74.3	64.4	37.0	76.6	60.9	6.6	−0.9	92.7	87.5	51.0	31.8	78.5	63.3
6220	5	28.6	−3.7	−3.7	9.1	8.4	9.0	8.1	92.4	90.1	1.2	−0.1	9.8	8.5	9.0	8.0	93.4	93.4
	10	57.1	−7.5	−7.5	11.1	10.5	9.3	8.2	84.5	80.9	2.2	−0.4	10.2	8.7	8.7	7.8	91.7	91.8
	20	114.2	−15.0	−15.05	17.0	16.6	9.7	8.4	55.8	46.0	4.7	−1.1	11.6	9.2	8.1	7.4	85.8	87.7
9120	5	0.5	−0.2	0.1	59.5	21.8	66.7	1.4	90.6	2.3	7.1	0.4	67.4	24.2	72.1	1.4	90.5	2.3
	10	1.0	−0.1	0.1	59.5	21.9	66.7	1.4	90.6	2.3	13.3	0.6	73.2	25.6	69.6	1.3	90.3	2.3
	20	2.0	−0.1	0.1	59.5	21.8	66.7	1.4	90.6	2.3	28.4	1.1	89.0	29.2	61.7	1.2	89.1	2.3
9210	5	8.05	−0.4	−0.4	13.0	6.8	14.6	4.9	96.2	67.3	3.1	−0.2	14.1	7.0	12.2	4.8	91.6	67.7
	10	16.1	−0.8	−0.8	13.0	6.7	14.6	4.7	96.1	65.1	6.2	−0.3	15.4	7.1	11.7	4.7	88.5	65.8
	20	32.2	−1.6	−1.6	12.9	6.4	14.6	4.4	95.7	60.3	13.1	−0.7	19.6	7.1	10.6	4.3	75.0	61.8
92A0	5	23.6	−2.0	−2.0	10.8	7.6	9.6	6.9	90.5	89.2	2.9	0.1	12.6	8.0	10.6	7.1	92.2	91.4
	10	46.9	−4.1	−4.1	11.2	8.2	9.7	6.7	87.9	84.0	5.9	−0.3	14.0	8.1	10.2	6.9	89.2	89.8
	20	93.4	−4.3	−4.3	12.9	9.1	13.4	7.6	92.3	76.0	12.5	−1.3	20.4	9.6	11.9	7.3	80.0	80.2
9330	5	8.1	−0.7	−0.6	30.9	9.7	24.7	6.4	85.1	75.9	4.6	0.1	33.7	10.3	24.5	6.4	85.0	76.9
	10	116.3	−1.4	−1.3	30.8	9.4	24.9	6.2	84.7	73.4	8.9	−0.0	36.0	10.4	23.4	6.2	83.7	75.0
	20	32.6	−2.8	−2.7	30.8	9.1	25.1	5.8	84.0	68.0	19.3	−0.2	42.7	10.7	21.2	5.7	77.9	71.3
4210	5	5.2	−0.9	−0.9	10.6	8.1	13.8	7.3	97.9	84.2	2.4	−0.4	12.3	8.3	11.6	7.1	93.1	85.0
	10	10.3	−1.9	−1.9	10.7	8.1	13.8	7.2	97.5	81.1	4.4	−0.8	13.0	8.4	11.1	6.9	90.5	82.7
	20	20.5	−3.8	−3.8	11.1	8.4	13.9	6.8	96.2	73.8	8.9	−1.8	15.4	8.6	10.2	6.5	81.4	77.7
9420	5	6.8	−1.0	−1.0	12.5	8.8	13.3	8.1	95.0	85.8	3.1	−0.2	15.3	9.2	13.6	8.0	82.4	86.5
	10	13.6	−2.2	−2.2	12.6	8.8	13.3	7.9	94.2	82.9	6.0	−0.6	16.4	9.3	13.0	7.8	80.0	80.2
	20	27.2	−4.3	−4.3	12.9	9.1	13.4	7.6	92.3	76.0	12.5	−1.3	20.4	9.6	11.9	7.3	80.0	80.2

NRR = nonresponse rate (%), ENR = expected nonresponses in the sample (%), RB = relative bias (%), RRMSE = relative root mean squared error (%), ERSEE = expectation of relative standard error estimator (%), C95 = coverage of the 95% confidence interval (%), DE = double expansion estimator, DIF = difference estimator. Values in blue and green refer to DE and DIF estimation, respectively.

compliance with the mandatory reporting cycle of the HD (ex art. 17). This approach can be adapted to any other country needing to soundly estimate the area covered by a habitat under study, vegetation type or ecosystem that cannot be properly mapped.

The results of a simulation study showed that the design-based inference performed by means of two-phase sampling and the use of the DIF estimator exploiting previous knowledge of habitat presence, or its calibrated counterpart in the presence of nonresponse have the potential to improve precision with respect to the HT estimator achieved under SRSWOR, thus showing a considerable design effect. Adopting a small sampling fraction not greater than 0.04% of the survey area, the DIF estimator provides suitable precision with RSEs smaller than 15% if habitats are quite common (e.g., $M > 200$), if the HSSs are good proxies for habitat presence and if coverages are not smaller than 1% with respect to the survey areas. Therefore, the same general strategy can be efficiently applied to different habitat types, from grassland (e.g., 6220) to forests (e.g., 9210), simply by changing the set of environmental predictors. Moreover, these habitats are characterized by a large geographic distribution across the three Italian biogeographical regions (Cervellini et al., 2021), partially confirming the applicability of the strategy to different macroecological and biogeographical contexts. On the other hand, the strategy provides unsuitable precision when the portion of coverage is small with respect to the survey area (e.g., 5110)—a characteristic that reduces the precision of any sampling strategy in spatial surveys—when the survey area is too small (e.g., 9120) and when the correlation between HSSs and habitat presence is weak (e.g., 9330). Notably, however, these situations are likely to reduce the precision of most sampling strategies and not only that of our strategy.

From these results, it is also apparent that for sufficiently large coverages, the correlation between HSSs and habitat presence and the information on habitat presence that may be available from previous investigations and surveys are determinant factors to efficiently increase the precision of the DIF estimator or its calibrated counterpart. A less satisfactory issue concerns the RSE estimation and the construction of confidence intervals. However, it is well known that this issue is “slightly tricky” in spatial sampling (e.g., Grafström, 2012), when, as in our case, the selection of neighbouring units is avoided or reduced.

These findings are essential for producing reliable estimates of area coverage by each of the 124 habitat types in the study region for the forthcoming fifth reporting cycle (2019–2024). In addition, for each habitat, our sampling strategy and the related estimation of its precision allow for a statistically sound detection of changes and trends (Lengyel et al., 2018) in relation to future national reports (e.g., 6th report). Furthermore, in the context of a “mandated” monitoring (Lindenmayer and Likens, 2010), it is now possible to plan a rigorous program based on a sustainable effort for each reporting cycle (Lindenmayer et al., 2020).

5. Concluding remarks

The design we developed is based on two-phase sampling that ensures the geographic spread of sample units in the area covered by each habitat type, permitting the collection of information across the whole habitat, as well as the use of local clusters of sample units that allow the optimization of time effort spent moving among them. This design permits the production of sound statistical estimates of habitat coverage while simultaneously providing sound information about uncertainty. In addition, exploitation of the difference estimator permits us to positively include all the habitat occurrence data that are collected out of the sampling scheme proposed here in the resulting estimates, such as data from local surveys or management or monitoring plans within protected areas. While the assemblage of these data, collected at local scales without a probabilistic design, does not facilitate statistically sound inferences, available data are effectively exploited in this approach as auxiliary components, improving the quality of the estimates produced on the basis of only the data collected by the sampling scheme. Basically, this sampling strategy permits unification of the data collected under a well-designed probabilistic scheme with those opportunistically provided by all other available sources. Given the complex distribution of the various habitats, at the national, biogeographical and continental scales, this design-based approach permits the integration of a specifically defined and limited probabilistic sample with a likely larger sample lacking probabilistic features. Ultimately, this approach can be profitably used to arrange “mandated” monitoring plans at broad scales, such as the whole nation, biogeographical region or European Union, as is the case for the monitoring imposed by the Habitats Directive.

Data availability statement

Data will be made available upon request.

CRediT authorship contribution statement

Lorenzo Fattorini: Conceptualization. **Marco Cervellini:** Conceptualization. **Sara Franceschi:** Conceptualization, Data curation. **Michele Di Musciano:** Data curation. **Piero Zannini:** Data curation. **Alessandro Chiarucci:** Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Research contributing to this study was funded by the project “Development of a National Plan for Biodiversity Monitoring” (Italian National Institute for Environmental Protection and Research—ISPRA).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2022.109352.

References

Angelini, P., Casella, L., Grignetti, A. and Genovesi, P. (2016), Manuali per il monitoraggio di specie e habitat di interesse comunitario (Direttiva 92/43/CEE) in Italia: habitat, ISPRA, Serie Manuali e linee guida, 142/2016.

Álvarez-Martínez, J.M., Jiménez-Alfaro, B., Barquín, J., Ondiviela, B., Recio, M., Silió-Calzada, A., Juanes, J.A., 2018. Modelling the area of occupancy of habitat types with remote sensing. Methods Ecol. Evol. 9 (3), 580–593. https://doi.org/10.1111/2041-210X.12925.

Breidt, F.J., 1995. Markov chain designs for one-per-stratum sampling. Survey Methodology 21, 63–70.

Brooks, T.M., Pimm, S.L., Akçakaya, H.R., Buchanan, G.M., Butchart, S.H.M., Foden, W., Hilton-Taylor, C., Hoffmann, M., Jenkins, C.N., Joppa, L., Rondinini, C., 2019. Measuring terrestrial area of habitat (AOH) and its utility for the IUCN Red List. Trends Ecol. Evol. 34 (11), 977–986. https://doi.org/10.1016/j.tree.2019.06.009.

Brown, J.A., Robertson, B.L., McDonald, T., 2015. Spatially balanced sampling: application to environmental surveys. Procedia Environ. Sci. 27, 6–9. https://doi.org/10.1016/j.proenv.2015.07.108.

Carli, E., Massimi, M., Angelini, P., Casella, L., Attorre, F., Agrillo, E., 2020. How to improve the distribution maps of habitat types at national scale. Rendiconti Lincei. Scienze Fisiche e Naturali 31 (3), 881–888. https://doi.org/10.1007/s12210-020-00917-7.

CDR (2022). EIONET Central Data Repository. https://cdr.eionet.europa.eu/it/eu/art17/envxuwp6g/. Accessed on June 21st 2022.

Cervellini, M., Di Musciano, M., Zannini, P., Fattorini, S., Jiménez-Alfaro, B., Agrillo, E., Attorre, F., Angelini, P., Beierkuhnlein, C., Casella, L., Field, R., Fischer, J., Genovesi, P., Hoffmann, S., Irl, S.D.H., Nascimbene, J., Rocchini, D., Steinbauer, M., Vetaas, O.R., Chiarucci, A., 2021. Diversity of European habitat types is correlated with geography more than climate and human pressur. Ecol. Evol. https://doi.org/10.1002/ece3.8409.

Cervellini, M., Zannini, P., Di Musciano, M., Fattorini, S., Jiménez-Alfaro, B., Rocchini, D., Field, R., Vetaas, O.R., Irl, S.D.H., Beierkuhnlein, C., Fischer, J.C., Casella, L., Angelini, P., Genovesi, P., Nascimbene, J., Chiarucci, A., 2020. A grid-based map for the Biogeographical Regions of Europe. Biodiversity Data J. 8 https://doi.org/10.3897/BDJ.8.e53720.

Chiarucci, A., Di Biase, R.M., Fattorini, L., Marcheselli, M., Pisani, C., 2018. Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness. Ann. Appl. Stat. 12 (3), 1679–1699. https://doi.org/10.1214/17-AOAS1126.

CLC (2018), Corine Land Cover. Version is v.2020_20u1. https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download.

Coops, N.C., Wulder, M.A., 2019. Breaking the Habit (at). Trends Ecol. Evol. 34 (7), 585–587. https://doi.org/10.1016/j.tree.2019.04.013.

Cowles, J., Templeton, L., Battles, J.J., Edmunds, P.J., Carpenter, R.C., Carpenter, S.R., Paul Nelson, M., Cleavitt, N.L., Fahey, T.J., Groffman, P.M., Sullivan, J.H., Neel, M. C., Hansen, G.J.A., Hobbie, S., Holbrook, S.J., Kazanski, C.E., Seabloom, E.W., Schmitt, R.J., Stanley, E.H., Vander Zanden, J.M., 2021. Resilience: insights from the US LongTerm Ecological Research Network. Ecosphere 12 (5), e03434. https://doi.org/10.1002/ecs2.3434.

Davies, C.E., Moss, D., Hill and M.O. (2004), EUNIS habitat classification revised 2004, Report to: European Environment Agency-European Topic Centre on Nature Protection and Biodiversity, 127–143.

Delbosc, P., Lagrange, I., Rozo, C., Bensettiti, F., Bouzillé, J.B., Evans, D., Lalanne, A., Rapinel, S., Bioret, F., 2021. Assessing the conservation status of coastal habitats under Article 17 of the EU Habitats Directive. Biol. Conserv. 254, 108935 https://doi.org/10.1016/j.biocon.2020.108935.

DEM20, Digital Elevation Model. Rete del Sistema Informativo Nazionale Ambientale. SINAnet, https://www.sinanet.isprambiente.it/it/sia-ispra/download-mais/dem20/view.

Drakou, E.G., Kallimanis, A.S., Mazaris, A.D., Apostolopoulou, E., Pantis, J.D., 2011. Habitat type richness associations with environmental variables: a case study in the Greek Natura 2000 aquatic ecosystems. Biodivers. Conserv. 20 (5), 929–943. https://doi.org/10.1007/s10531-011-0005-4.

EIONET (2022), Article 17 web tool. Available from https://nature-art17.eionet.europa.eu/article17/habitat/summary/?period=3&group=.

Ellwanger, G., Runge, S., Wagner, M., Ackermann, W., Neukirchen, M., Frederking, W., Müller, C., Ssymank, A., Sukopp, U., 2018. Current status of habitat monitoring in the European Union according to Article 17 of the Habitats Directive, with an emphasis on habitat structure and functions and on Germany. Nature Conservation 29 (October), 57–78. https://doi.org/10.3897/natureconservation.29.27273.

European Commission (2020). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. EU Biodiversity Strategy for 2030 - Bringing nature back into our lives.

European Commission - DG environment (2021). https://ec.europa.eu/environment/nature/natura2000/index_en.htm. Accessed July 2021 the 29th.

European Enviroment Agency (2013), EEA Reference Grid. Available from https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2.

Fahrig, L., 2013. Rethinking patch size and isolation effects: the habitat amount hypothesis. J. Biogeogr. 40 (9), 1649–1663. https://doi.org/10.1111/jbi.12130.

Fattorini, L., Franceschi, S., Maffei, D., 2013. Design-based treatment of unit nonresponse in environmental surveys using calibration weighting. Biometrical J. 55 (6), 925–943. https://doi.org/10.1002/bimj.201100262.

Foster, S., 2020. DMBHdesign: An R-package for efficient spatial survey designs. Methods Ecol. Evol. 12 (3), 415–420. https://doi.org/10.1111/2041-210X.13535.

Grafström, A., 2012. Spatially correlated Poisson sampling. J. Stat. Planning Inference 142 (1), 139–147. https://doi.org/10.1016/j.jspi.2011.07.003.

Grafström, A., Lundström, N.L.P., 2013. Why well spread probability samples are balanced. Open J. Stat. 3 (1), 36–41. https://doi.org/10.4236/ojs.2013.31005.

Hall, L.S., Krausman, P.R., Morrison, M.L., 1997. The Habitat Concept and a Plea for Standard Terminology. Wildlife Society Bull. 25 (1), 173–182.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. J. R. Stat. Soc. Series C (Appl. Stat.) 28 (1), 100–108. https://doi.org/10.2307/2346830.

Haziza, D., Thompson, K.J., Yung, W., 2010. The effect of nonresponse adjustments on variance estimation. Survey Methodol. 36 (1), 35–43.

ISTAT (2022), Confini delle unitá amministrative a fini statistici, https://www.istat.it/it/archivio/222527. Accessed on June 21st 2022.

ISTAT (2020), Sezioni di Censimento Litoranee. Shape file della linea litoranea, https://www.istat.it/it/archivio/137341. Accessed on March 2010.

Keyes, A.A., McLaughlin, J.P., Barner, A.K., Dee, L.E., 2021. An ecological network approach to predict ecosystem service vulnerability to species losses. Nature Commun. 12 (1), 1–11. https://doi.org/10.1038/s41467-021-21824-x.

Lengyel, S., Kosztyi, B., Schmeller, D.S., Henry, P.Y., Kotarac, M., Lin, Y.P., Henle, K., 2018. Evaluating and benchmarking biodiversity monitoring: Metadata-based indicators for sampling design, sampling effort and data analysis. Ecol. Ind. 85, 624–633. https://doi.org/10.1016/j.ecolind.2017.11.012.

Lindenmayer, D.B., Likens, G.E., 2009. Adaptive monitoring: a new paradigm for long-term research and monitoring. Trends Ecol. Evol. 24 (9), 482–486. https://doi.org/10.1016/j.tree.2009.03.005.

Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143 (6), 1317–1328. https://doi.org/10.1016/j.biocon.2010.02.013.

Lindenmayer, D.B., Likens, G.E., 2018. Maintaining the culture of ecology. Front. Ecol. Environ. 16 (4) https://doi.org/10.1002/fee.1801, 195–195.

Lindenmayer, D.B., Woinarski, J., Legge, S., Southwell, D., Lavery, T., Robinson, N., Scheele, B., Wintle, B., 2020. A checklist of attributes for effective monitoring of threatened species and threatened ecosystems. J. Environ. Manage. 262, 110312 https://doi.org/10.1016/j.jenvman.2020.110312.

Martinez-Harms, M.J., Bryan, A., Balvanera, P., Law, E.A., Rhodes, J.R., Possingham, H. P., Wilson, K.A., 2015. Making decisions for managing ecosystem services. Biol. Conserv. 184, 229–238. https://doi.org/10.1016/j.biocon.2015.01.024.

Mitchell, S.C., 2005. How useful is the concept of habitat?–A critique. Oikos 110 (3), 634–638. https://doi.org/10.1111/j.0030-1299.2005.13810.x.

MiTE (2021), National cartography of the network of sites. https://www.mite.gov.it/pagina/cartografie-rete-natura-2000-e-aree-protette-progetto-natura.

Mulder, C., Bennett, E.M., Bohan, D.A., Bonkowski, M., Carpenter, S.R., Chalmers, R., Cramer, W., Durance, I., Eisenhauer, N., Fontaine, C., Haughton, A.J., Hettelingh, J. P., Hines, J., Ibanez, S., Jeppesen, E., Krumins, J.A., Ma, A., Mancinelli, G., Massol, F., Woodward, G., 2015. 10 years later: revisiting priorities for science and society a decade after the millennium ecosystem assessment. Adv. Ecol. Res. 53, 1–53. https://doi.org/10.1016/bs.aecr.2015.10.005.

R Core Team (2020), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Sampford, M.R., 1967. On sampling without replacement with unequal probabilities of selection. Biometrika 54 (3–4), 499–513. https://doi.org/10.2307/2335041.

Särndal, C.E., Swensson, B., Wretman, J., 1992. Model assisted survey sampling. Springer Science & Business Media.

Smith, R.J., Gray, A.N., 2021. Strategic monitoring informs wilderness management and socioecological benefits. Conservation Sci. Practice 3 (9), e482. https://doi.org/10.1111/csp2.482.

Wolter, K.M., 2007. Introduction to variance estimation, (2nd ed),. Springer, p. 53.

SYapp, R.H., 1922. The concept of habitat. J. Ecol. 10 (1), 1–17.