

# Complex national sampling design for long-term monitoring of protected dry grasslands in Switzerland

Yves Tillé · Klaus Ecker

Received: 29 December 2012 / Revised: 13 June 2013 / Published online: 2 August 2013  
© Springer Science+Business Media New York 2013

**Abstract** We describe a probabilistic sampling design of circular permanent plots for the long-term monitoring of protected dry grasslands in Switzerland. The population under study is defined by the perimeter of a national inventory. The monitoring focus is on the species composition of the protected grassland vegetation and derived conservation values. Efficient trend estimations are required for the whole country and for some predefined target groups (six biogeographical regions and eleven vegetation types). The target groups are equally important regardless of their size. Consequently, intensified sampling of the less frequent groups is essential for sample efficiency. The prior information needed to draw a targeted sample is obtained from the sampling frame and external databases. The logistics and generalized delineation of the target population may pose further problems. Thus, investments in fieldwork and travel time should be well balanced by selecting a cluster sample. Second, any access problems in the field and non-target units in the sample should be compensated for by selecting reserve plots as they otherwise may considerably reduce the effective sample size. Finally, the design has to be flexible as the sampling frame may change over time and sampling intensity might have to be adjusted to redefined budgets or requirements. Likewise, the variables and biological items of interest may change. To fulfil all these constraints and to optimally use the available prior information, we propose a multi-

---

Handling Editor: Pierre Dutilleul.

---

Y. Tillé (✉)

Institute of Statistics, Faculty of Economics, University of Neuchâtel, Pierre à Mazel 7,  
2000 Neuchâtel, Switzerland  
e-mail: yves.tille@unine.ch

K. Ecker

Ecosystem Dynamics, Biodiversity and Conservation Biology, Swiss Federal Institute for Forest,  
Snow and Landscape Research WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland  
e-mail: klaus.ecker@wsl.ch

stage self-weighted unequal probability sampling design. The design uses modern techniques such as: balanced sampling, spreading, stratified balancing, calibration, unequal probability sampling and power allocation. This sampling design meets the numerous requirements of this study and provides a very efficient estimator.

**Keywords** Balanced sampling · National scale · Two-stage sampling · Sampling design · Unequal probability · Vegetation

## 1 Introduction

Biodiversity conservation is one of the greatest challenges of the 21st century since biodiversity has declined at unprecedented rates in recent decades. As a consequence, the international community has made various commitments to significantly reduce the loss of biodiversity. Starting with the ‘Convention on Biological Diversity’ of Rio in 1992, the international community then declared targets to halt biodiversity loss at the Johannesburg Summit in 2002 and reinforced the issue in 2010 with the Aichi Biodiversity targets for 2020. The various treaties express the commitment of the signing countries to take responsibility for their individual biological heritage.

In densely populated Switzerland, the habitats requiring conservation measures include man-made habitats in the cultural landscape such as dry grasslands. These habitats are considered valuable because they host a high biological diversity with many rare and threatened species. For example, 35 % of the vascular plants (Moser et al. 2002) and 50 % of some animal groups in dry meadows are Red List species (Eggenberg et al. 2001). However, about 95 % of these grasslands vanished last century (Lachat et al. 2010). The remaining sites are endangered by improper management, abandonment, nutrient inputs and habitat fragmentation. To record the remaining dry grasslands in Switzerland, the Swiss Federal Office for the Environment (FOEN) started an inventory in 1995. The main criterion for selecting a site to be listed and mapped was the occurrence and minimum extent of some typical dry grassland vegetation types. The most valuable sites (around 3,000 sites or ‘objects’), which altogether make up an area of approximately 21,000 ha or 1–2 % of the agricultural landscape, were designated for national protection in 2010. To maintain the conservation value of the sites, farmers receive subsidies for proper conservation management, which usually involves either mowing late in the year or extensive grazing by cattle, sheep, goats or donkeys.

The conservation values of a habitat may nevertheless deteriorate despite protection and financial compensation for adaptive management (Klaus 2007; Bergamini et al. 2009). To obtain reliable information on the development of conservation values in the protected sites, FOEN initiated a national and long-term (i.e. one decade onwards) monitoring programme in 2011 (Bergamini and Holderegger 2012). Since the conservation status of a site was determined by the occurrence of some typical dry grassland vegetation types, the main focus for monitoring was placed on plant species composition, which indicates the prevailing environmental conditions, is sensitive to the management system and can be used to quantify conservation values. Species

composition is therefore a widely used target item in monitoring semi-natural habitats (Kapfer et al. 2011; Koch et al. 2011; Thompson et al. 2011). The species composition data is collected in the field on a sample of 10 m<sup>2</sup> permanent plots, which are each revisited every 6 years. All vascular plants on each plot are listed. In addition, coverage estimates are made and various attributes describing the environment of the plot are recorded.

The collected data provide a range of variables or indicators for detecting ecological changes and changes in conservation values. For example, species or community turnover may be analysed based on abundance data (occurrence, plant frequency, plant cover), biodiversity indices, similarities or classifications (Legendre and Legendre 1998). Another focus is on community properties which describe the dominant habitat conditions. The concept of plant indicator values is the most prominent example. These values are assigned to vascular plants based on expert knowledge (Landolt et al. 2010; Ellenberg et al. 1992). The values may be numerically recalibrated using large data sets (Feldmeyer-Christe et al. 2007). The individual values rank the species according to their realized optima along various gradients such as soil pH, water supply, light supply, soil fertility, humus content and temperature. Averaged within a plot, they have proven to be a reliable substitute for more cost intensive geo-physical and geo-chemical field measurements (Diekmann 2003). Thus, not surprisingly, they have been widely used in plant ecology (Odland 2009) and vegetation monitoring (Graf et al. 2010).

The species data recorded on the plots allow a multitude of indicators to be derived and analysed. As a result, they may be useful for addressing new questions that arise, e.g. due to new priorities in environmental policy. Instead of optimizing the survey for existing risk scenarios, we less specifically organize the sampling to obtain a sufficient representation of the predefined vegetation types and some biogeographical regions. More specifically, at the planning stage, the monitoring is intended to accommodate 6 regions and 11 vegetation types. The subpopulations are of equal interest regardless of their extent, while estimations of the total population (the whole inventory) are of overall importance.

The choice of the appropriate sampling strategy is crucial for conducting a valid and efficient inference. In ecology, non-probabilistic sampling designs are commonly applied to assess the relationships between a response and a set of predictor variables over time and space (Albert et al. 2010). The corresponding methodology used for estimation is based on model assumptions. Consequently, the estimates refer to a theoretical infinite superpopulation (Gruijter et al. 2006) and are highly dependent on the correct model specification (Särndal et al. 1992). This strategy is inadequate for monitoring a national conservation network, where totals, means, proportions and their variances must be reliably estimated of the finite conservation area. The classical sampling approach in this setting combines probabilistic sampling and design-based statistics. This strategy provides model-free estimates of the finite study population, but relies on known inclusion probabilities of the sampling units and requires estimators that reflect the underlying sampling design.

The sampling design determines the efficiency of the survey. In theory, equal probability designs, i.e. simple random or random systematic designs, are to be favoured because they allow flexible and simple analyses. Such designs would be optimal to

accommodate periodic frame updates, e.g. additions of sites to the inventory, and sample restructuring, e.g. sample enlargement or sample reduction in specific areas (Overton and Stehman 1996). Equal probability designs are, however, associated with high travel expenses when the sampling population is scattered over a wide area. In large scale surveys the sample units should be clustered to reduce the travel costs. Equal probability designs are also less suitable, when quantities of rare vegetation types are to be estimated. In this case we aim to preferentially sample the less frequent types, which requires using unequal inclusion probabilities based on prior knowledge about them.

Sample optimization involves controlling the sampling variance. This variance is the expected sample-to-sample variation of an estimate under a given sampling design. The efficiency of a sample is improved when the sampling variance is minimized for the predefined target variables, e.g. the total area of a target vegetation type. The solution is to stratify or balance the sample on auxiliary variables which are correlated to the variables of interest. For the same reason the sample units and clusters should be well distributed in space and time. The spatial spreading is desired to avoid collecting redundant information from close sites and dates. Another issue to be considered in the sampling process is how to deal with plots that yield nonresponse or non-target response in the field. Sample units may be located in non-target vegetation due to cartographic generalization of the study perimeter. Or sample units may yield nonresponse due to access problems in the field. Both situations may considerably reduce the effective sample size and should be compensated for by selecting reserve plots.

In this paper, we outline a balanced multi-stage sampling design which fulfils all budget constraints and fieldwork logistics, while ensuring sample efficiency for rare subpopulations and still remaining a probabilistic sample. The design includes sampling techniques such as balanced sampling, spreading, stratified balancing, calibration, unequal probability sampling and power allocation. The aim of the complex design is to provide unbiased and accurate estimates. The paper is structured as follows. In Sect. 2, the sampling frame is described and the notation is defined. The aim of the sampling design is presented in Sect. 3. In Sect. 4, we describe the computation of a rareness index. The index is built to over-represent at the same time the small regions and the rare vegetation types. Hence, the rareness index is used to compute the inclusion probabilities in Sect. 5. The sampling design is described in Sect. 6. The way of estimating a total is defined in Sect. 7 and a variance estimator is given in Sect. 8. In Sect. 9, we close with a short discussion.

## 2 Statistical units and auxiliary information

### 2.1 The inventory of dry grasslands in Switzerland

The sampling frame is given by the database of the Swiss inventory of dry grasslands of national importance. This cartography of dry grassland polygons is the result of a nationwide field campaign, conducted between 1995 and 2006. Data collection and

data evaluation was standardized according to [Eggenberg et al. \(2001\)](#). The central part of the field campaign was the mapping of dry grasslands based on a modular vegetation key. For a grassland patch to be mapped as a polygon, the patch had to satisfy a set of requirements including size and quality. To comply with the fine-scale vegetation mosaic, a polygon was generally defined as a complex of contiguous vegetation groups. The delineation rules allowed non-target units such as bosks or anthropogenic structures such as houses to be enclosed in a polygon. For operational reasons, contiguous polygons were formally organized into grassland objects. The vegetation groups recorded in the field were a posterior aggregated into 18 main vegetation types, and the proportions of the new categories within a polygon and object were computed accordingly. The new classification was used to define the conservation value of an object and thus to decide if an object was of national importance or not. The inventory gained official status in February 2010. The corresponding database describes 10,059 polygons, varying in size from 0.05 to 67 ha (mean: 2.1 ha), in 2939 dry grassland objects (3.4. polygons per site), covering in total an area of 214 km<sup>2</sup> (0.5 % of Switzerland). The enclosure of non-target units amounts to 49 km<sup>2</sup> or 23 % of the sample frame. The inventory was revised in December 2011 after the selection of the sample. For sampling with unequal intensities, we aggregated the rarest vegetation types to obtain 11 target vegetation types. The rationale was to define target groups that result minimum subsample rates of 80 units without putting too much stress on the inclusion probabilities.

## 2.2 External information

Temperature is the main bioclimatic factor controlling the distribution pattern of vegetation in Switzerland. It is highly correlated to altitude, but it is more directly linked to the physiology of plants. To use this strong relationship for the sampling, we employed the variable ‘annual degree days of the growing season’, which considers 0 °C as the minimum temperature for plant growth ([Zimmermann and Kienast 1999](#)). The variable is derived from temperature maps spatially interpolated from long-term meteorological recordings using a digital terrain model.

## 2.3 Objects, polygons and plots

We aimed to select  $n=2,100$  points as the centres of circular 10 m<sup>2</sup> dry grassland plots, i.e. with a radius of 1.78 m. The plots are distributed within  $m = 420$  objects of the inventory, so that the sampling design has two stages. The objects are the primary units (PU), and the plots the secondary units (SU). Using a clustered sample helps to save travel costs. The degree of clustering is based on the experience from other monitoring programs. The resulting sample size depends on the available budget.

The objects in the original database vary greatly in size. So we split the largest objects into smaller ones, which increased the number of objects to  $M=3,085$  units. [Figure 1](#) shows an example of a selected object and the distribution of the sample plots across the polygons.



**Fig. 1** Example of a dry grassland object in the Swiss Jura mountains containing 9 polygons (*black delineation*). Six sample plots (*white crosses*) and two reserve plots (*white circles*, see Sect. 6.5) were selected for recording in the field. Data source orthophoto: swissimage © 2012 swisstopo (DV 033594)

## 2.4 Notation

We refer to the following notation. The PU are the objects denoted by:

- object  $O_i$ ,  $i = 1, \dots, M$ ,
- area of the object  $A(O_i)$  in  $\text{m}^2$ ,
- inclusion probability of an object  $\pi_{1,i}$ ,
- rareness index of the object  $R(O_i)$  (higher values denote rarer and more valuable objects),
- sample  $S_1$  of objects.

The rareness index or ‘rareness’ is a measure used to compute the final inclusion probabilities of the circular plots. The aim is to over-represent the rare types of grasslands and to under represent the more common ones. The computation of the rareness index is described in Sect. 4.

The objects are composed of a set of contiguous polygons.

- polygon  $P_k$ ,  $k = 1, \dots, N_i$ , with  $k \in O_i$ ,
- number  $N_i$  of polygons in  $O_i$ ,
- area of the polygon  $A(P_k)$  in  $\text{m}^2$ ,
- rareness of the polygon  $R(P_k)$  (higher values denote rarer and more valuable polygons).

Note that the rareness index is supposed to be linear such that

$$R(O_i) = \frac{\sum_{k \in O_i} A(P_k)R(P_k)}{A(O_i)}. \tag{1}$$

The linearity of the rareness index is used below to determine the inclusion probabilities. If rareness is determined at the level of the polygon, it can thus be computed at the level of the objects by using (1).

Circular plots of 10 m<sup>2</sup> are selected in the polygons. The following notation is used:

- $Q_r$  denotes a circular plot in a polygon,
- $A(Q_r) = A(Q) = 10 \text{ m}^2$  is the area of plot  $Q_r$ , the area is the same for all the plots,
- $\pi_r$  is the inclusion probability of this plot,
- $y_r$  is the variable of interest measured in the circular plot, for instance, the area occupied by a plant,
- $S_{2,k}$  denotes the sample of the circular plots selected in polygon  $P_k$ .

We assume that the polygons can be partitioned into circular plots. This is obviously mathematically impossible but without practical consequence for the computation of the estimators. Under this simplifying hypothesis, the number of sampling units (i.e circular plots) in a polygon equals  $A(P_k)/A(Q)$ . This quantity is not integer. Again, this is without consequence for the reasoning. Our aim is to select circular plots with inclusion probabilities proportional to the rareness index. More specifically, we would like to select plots  $Q_r$  with a probability  $\pi_r$  proportional to the rareness  $R(P_k)$  of the polygon  $P_k$  to which  $Q_r$  belongs.

### 3 Objectives of the sampling designs

The objectives of the sampling design are to:

- apply a two-stage sampling by selecting a sample of  $m = 420$  objects as PU to decrease the travel costs of the surveyors,
- spread as well as possible the sample of objects within Switzerland,
- select a sample of  $n = 2,100$  circular plots with inclusion probabilities proportional to the rareness index of the polygon to which each plot belongs,
- balance as well as possible the sample on a number of known auxiliary variables.

When units are selected with unequal probabilities, the Horvitz and Thompson (1952) estimator of a total consists of weighting each unit by the inverse of its inclusion probability. This estimator is unbiased. When several auxiliary variables are available for each unit of the population, a balanced sample can be selected in the sense that the Horvitz–Thompson estimator is equal or almost equal to the total of these auxiliary variables (see the definition given in Deville and Tillé 2004; Tillé 2006, p. 148). The way of balancing the grassland sample is described in Sect. 6.3.

‘Spatial spreading’ means here trying to avoid the selection of two objects or plots that are very close geographically (see the definition in Grafström and Tillé 2013). Indeed, selecting neighbouring objects or plots is inefficient if spatial autocorrelation

exists. The auxiliary information provided from the inventory of dry grasslands and external databases is known at the level of the polygons and, consequently, of the objects, but not at the level of the plots.

#### 4 Computation of the rareness index

The polygons can be assigned to biogeographical regions, which can then be used for stratification. Similarly, the polygons can be grouped according to their dominant vegetation type. The sample is used to produce estimations with respect to these two stratifications. Since estimations within biogeographical regions and vegetation types are nearly as important as the national estimation, we oversample the small and rare strata. The rareness index is thus constructed in order to oversample the rare vegetation types and those regions with small proportions of the sampling population.

The construction uses the method of power allocation (Bankier 1988) to find a compromise between two allocations:

- *Proportional allocation*, where the sample size is proportional to the population stratum size, and
- *Equal sample size*, where the same sample size is used in each stratum.

Power allocation is governed by a parameter  $\alpha \in [0, 1]$  that enables weighting of the importance of the domain estimation (i.e. the estimation of the predefined target groups) and the overall estimation. If  $\alpha = 1$ , the overall estimation is favoured. If  $\alpha = 0$ , the domain estimation is favoured and the same sample size is taken in each domain regardless of the size of this domain in the population. We tested several values for  $\alpha$ , and finally decided to favour the domain estimation by choosing a value  $\alpha = 0.1$ . Indeed, the aim of this survey is to pay a particular attention to the rare vegetation types and the small regions.

The importance of each domain in the sample is computed from the areas of polygons by means of a power allocation as follows. Suppose that  $U_1, \dots, U_h, \dots, U_H$  are the strata of the regions and that  $V_1, \dots, V_g, \dots, V_G$  are the strata of vegetation types. Let  $A(U_h)$  be the area of a region stratum  $U_h$  in the population, and  $a_{h,U}$  the area of the same stratum in the sample. A power allocation then determines  $a_{h,U}$  by

$$a_{h,U} = \frac{A^\alpha(U_h)}{\sum_{\ell=1}^H A^\alpha(U_\ell)}.$$

For the vegetation types, we can apply the same method. Let  $A(V_g)$  be the area of a vegetation type stratum  $V_g$  in the population, and  $a_{g,V}$  the area of the same stratum in the sample. A power allocation then determines  $a_{g,V}$  by

$$a_{g,V} = \frac{A^\alpha(V_g)}{\sum_{\ell=1}^G A^\alpha(V_\ell)}$$

Moreover, in each polygon, variable REST measures the area of non-target vegetation in the polygon without specific conservation value. As we would like to limit the area



of REST in the sample, we define an index of rareness  $R(P_k)$  for the polygons that satisfies the power allocation for both stratifications simultaneously.

$$\sum_{k \in U_h} \frac{A(P_k)}{A} R(P_k) = a_{h,U}, \text{ for all the strata } U_1, \dots, U_h, \dots, U_H, \tag{2}$$

$$\sum_{k \in V_g} \frac{A(P_k)}{A} R(P_k) = a_{g,V}, \text{ for all the strata } V_1, \dots, V_g, \dots, V_G, \tag{3}$$

where  $A$  is the total area. Moreover, we would like to reduce the area of the sample that is in the REST category.

$$\sum_{i=1}^M \sum_{k \in O_i} R(P_k) \frac{REST_k}{REST} = 0.8, \tag{4}$$

where  $REST$  is the sum of the variable REST for the entire population. A moderate reduction value of 0.8 was chosen to keep the dispersion of the inclusion probabilities low.

A rareness index  $R(P_k)$  that satisfies the set the constraints given in (2), (3) and (4) can be determined using the calibration method proposed by [Deville and Särndal \(1992\)](#). The calibration function used here is the logistic function with imposed bounds (0.38, 5.9) to limit the dispersion of the rareness index. The method starts with equal weights, which are then calibrated on the constraints (2), (3) and (4) to obtain the rareness index. This index is large at the same time for the small regions and the rare vegetation types. We then construct a sampling design to select circular plots with inclusion probabilities proportional to the rareness index.

## 5 Inclusion probabilities

### 5.1 General principles

The sampling design is a self-weighted two-stage design ([Särndal et al. 1992](#), pp. 132–154), with an additional contrast on the inclusion probabilities that depends on the rareness of the polygons. The final requirement of the sampling design is that the final inclusion probabilities  $\pi_r$  of the plots must be proportional to the rareness index  $R(P_k)$  of the polygon  $P_k$  to which  $Q_r$  belongs. The objects are accordingly selected with inclusion probabilities proportional to the product of their area and their rareness. The number of selected circular plots is generally the same in all the objects (except for the rare-large ones), which makes it easier to organize work activities. Hence, the self-weighting allows us to gain control of the survey effort per object or site. Let  $y_r$  be the variable of interest measured in circular plot  $Q_r$ . The total of  $y_r$  in the sampling frame can then be estimated by weighting the selected units by the inverse of their inclusion probabilities ([Horvitz and Thompson 1952](#)). The estimation procedure is described precisely in Sect. 7.

### 5.2 Definition of the inclusion probabilities of the objects

At the first stage, we select a sample of  $m = 420$  objects with unequal probabilities proportional to the product of the rareness and the area  $A(O_i)R(O_i)$ . Thus:

$$\pi_{1,i} = \min [c A(O_i) R(O_i), 1]$$

where  $c$  is defined such that

$$\sum_{i=1}^M \min [c A(O_i) R(O_i), 1] = m.$$

The determination of the value of  $c$  can be obtained in a simple algorithm defined, for instance, in Tillé (2006, pp. 18–19) and implemented in the R sampling package (Tillé and Matei 2012). Note that the large rare objects can have an inclusion probability equal to 1. In this case, they are always selected in the sample.

### 5.3 Number of plots in each object

Next, we distribute the predefined total sample size of  $n=2,100$  circular plots within the selected objects by computing  $n_i$  the expected number of circular plots to select in object  $O_i$ . The final requirement is that the total sample size  $n$  is given and that the plots must be selected with an inclusion probability proportional to their rareness, i.e. to the rareness of the polygon to which they belong.

In object  $O_i$ , the number of circular plots is computed as follows:

$$n_i \propto \frac{A(O_i)R(O_i)}{A(Q)\pi_{1,i}} = \frac{A(O_i)R(O_i)}{A(Q) \min [c A(O_i)R(O_i), 1]}.$$

Thus,  $n_i$  can be computed by

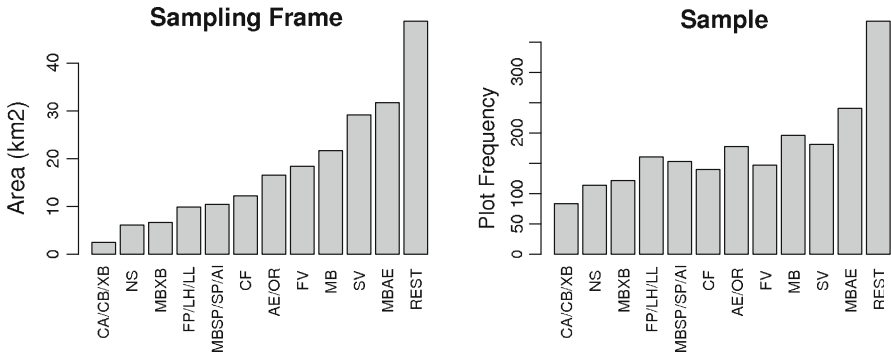
$$n_i = n \frac{\frac{A(O_i)R(O_i)}{\pi_{1,i}}}{\sum_{j \in S_1} \frac{A(O_j)R(O_j)}{\pi_{1,j}}} = n \frac{A(O_i)R(O_i)}{\pi_{1,i} \sum_{j=1}^M A(O_j)R(O_j)},$$

where  $S_1$  is the sample of objects. Indeed,

$$\sum_{j \in S_1} \frac{A(O_j)R(O_j)}{\pi_{1,j}} = \sum_{j=1}^M A(O_j)R(O_j).$$

Two cases must be distinguished:

- When  $\pi_{1,i} < 1$ , then  $cA(O_i)R(O_i) < 1$ . In this case,  $n_i \propto c^{-1}$ .
- When  $\pi_{1,i} = 1$ , then  $cA(O_i)R(O_i) \geq 1$ . In this case,  $n_i \propto A(O_i)R(O_i)$ .



**Fig. 2** Distribution of the protected dry grassland across 11 target vegetation types and the non-target enclosures (= REST). Some vegetation types are grouped since they were too small in size to be efficiently sampled. The figure on the left indicates the area of each vegetation type in the sampling frame. The figure on the right contains the number of selected circular plots for each vegetation type. CA Caricion austroalpinae, CB Cirσιο-Brachypodion, FP Festucion paniculatae, XB Xerobromion, LL low diversity–low altitude, MBSP Transition Mesobromion/Stipo-Poion, OR Origanetalia, AI Agropyron intermedii, SP Stipo-Poion, NS Nardion strictae, MBXB Transition Mesobromion/Xerobromion, LH low diversity–high altitude, CF Caricion ferrugineae, AE Arrhenatherion elatioris, FV Festucion varia, MB Mesobromion, SV Seslerion varia, MBAE Transition Mesobromion/Arrhenatherion

The same expected number of circular plots are selected in each small non-rare object, while the expected number of plots in the large-rare objects is proportional to the product of the size of the object and the rareness. Notice also that  $n_i$  is generally non-integer.

Figure 2 shows the distribution of the area of each vegetation type in the sampling frame and the number of plots selected in each vegetation type. The rare vegetation types are overrepresented. Figure 3 contains the distribution of the area of each biogeographical region in the sampling frame and the number of plots selected in each region. The regions with smaller area of protected dry grassland area are overrepresented. This double property of overrepresentation is due to the use of the rareness index in the sampling design.

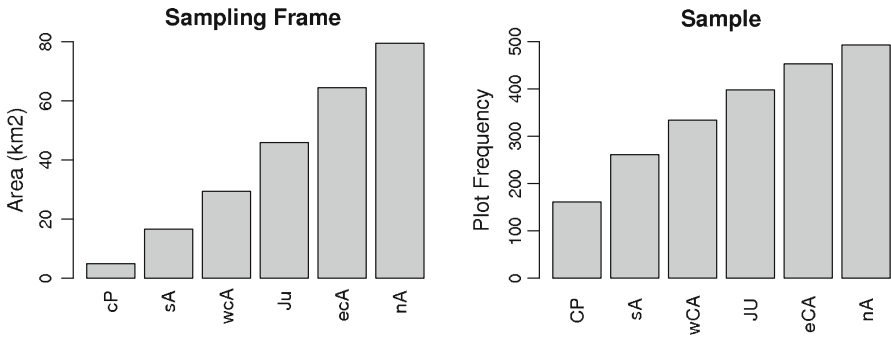
### 5.4 Number of circular plots in each polygon

The expected number of circular plots  $n_i$  in an object must be split into the polygons. Again, we over-represent the large-rare polygons. Let  $n_{k,i}$  denote the expected number of plots selected in polygon  $P_k$  that belongs to object  $O_i$

$$n_{k,i} = \frac{n_i A(P_k) R(P_k)}{A(O_i) R(O_i)}. \tag{5}$$

Thus

$$n_i = \sum_{k \in O_i} n_{k,i}.$$



**Fig. 3** Distribution across six target biogeographical regions. The figure on the left contains the area of each biogeographical region in the sampling frame. The figure on the right contains the number of selected circular plots for each biogeographical region. Regions with less protected dry grassland are overrepresented. CP central Plateau, sA southern Alps, wCA western Central Alps, JU Jura, eCA eastern Central Alps, nA northern Alps

Since  $n_{k,i}$  is not necessarily integer, it must be rounded. The best way is to round the  $n_{k,i}$  randomly in order to exactly satisfy the inclusion probabilities.

### 5.5 Summary of the inclusion probabilities

To summarize, the inclusion probabilities are:

- the inclusion probability of an object is given by

$$\pi_{1,i} = \min [cA(O_i)R(O_i), 1],$$

- the inclusion probability of a circular plot given that it belongs to polygon  $P_k$  is equal to the number of selected circular plots divided by the number of circular plots contained in this polygon, i.e.

$$\pi_{2,r|k} = \frac{n_{k,i}A(Q)}{A(P_k)}.$$

- the inclusion probability for the complete sampling design is thus

$$\begin{aligned} \pi_r &= \pi_{1,i}\pi_{2,r|k} = \pi_{1,i} \frac{n_{k,i}A(Q)}{A(P_k)} = \pi_{1,i}A(Q) \frac{\frac{n_i A(P_k)R(P_k)}{A(O_i)R(O_i)}}{A(P_k)} = \pi_{1,i}A(Q) \frac{n_i R(P_k)}{A(O_i)R(O_i)} \\ &= \pi_{1,i}A(Q) \frac{R(P_k)}{A(O_i)R(O_i)} n \frac{\frac{A(O_i)R(O_i)}{\pi_{1,i}}}{\sum_{j=1}^M A(O_j)R(O_j)} = \frac{nR(P_k)A(Q)}{\sum_{j=1}^M A(O_j)R(O_j)}. \end{aligned}$$

We thus obtain

$$\begin{aligned} \sum_{i=1}^M \sum_{k \in O_i} \sum_{r \in P_k} \pi_r &= \sum_{i=1}^M \sum_{k \in O_i} \sum_{r \in P_k} \frac{nR(P_k)A(Q)}{\sum_{j=1}^M A(O_j)R(O_j)} \\ &= \sum_{i=1}^M \sum_{k \in O_i} \frac{nA(P_k)R(P_k)}{\sum_{j=1}^M A(O_j)R(O_j)} = \sum_{i=1}^M \frac{nA(O_i)R(O_i)}{\sum_{j=1}^M A(O_j)R(O_j)} = n. \end{aligned}$$

## 6 The sampling design

### 6.1 Description of the sampling design

The sampling design consists of three steps:

1. selecting the objects,
2. rounding  $n_{k,i}$  to an integer number,
3. selecting a fixed number of circular plots in each polygon.

### 6.2 First step: selection of objects

We select a sample of objects with inclusion probabilities  $\pi_{1,i}$  that is balanced on the variables contained in vector  $\mathbf{x}_i$ , where

$$\mathbf{x}_i = \left( \pi_{1,i}, A(O_i), A(O_i) \times R(O_i), D_i, D_i^2, \right. \\ \left. \text{and all the variables with the area occupied by each type of vegetation} \right),$$

where variable  $D_i$  is the annual degree-days of the growing season (basis 0 °C) (Zimmermann and Kienast 1999). Variable  $D_i^2$  is the square of  $D_i$ . The bioclimatic variable is added to balance the sample on the main bioclimatic gradient in Switzerland. The quadratic term reflects the strong unimodal response curve of vegetation to this steep gradient. The sample is selected by using the doubly balanced spatial sampling design proposed by Grafström and Tillé (2013), which combines the cube method of Deville and Tillé (2004) and the spatial pivotal method of Grafström et al. (2012). This method ensures that the random sample of objects  $S_1$  is balanced on the balancing variables  $\mathbf{x}_i$ , such that

$$\sum_{i=1}^M \mathbf{x}_i \approx \sum_{i \in S_1} \frac{\mathbf{x}_i}{\pi_{1,i}}.$$

The cube method (Deville and Tillé 2004) and the Grafström and Tillé (2013) method are algorithms that are made of two phases: the flight phase and the landing phase. During the flight phase, the balancing equations are always exactly satisfied. The landing phase is necessary when it is not possible to find an exactly balanced

sample. This problem is referred as the rounding problem and is fixed by the landing phase. To improve the efficiency of the sampling design, the flight phase of the doubly balanced spatial sampling design is applied separately in each biogeographical region. The method is then applied again to the rounding problem as proposed by Chauvet (2009) to produce a sample that is almost balanced in each region and that is stratified by region. The map of Switzerland in Fig. 4 shows the distribution of dry grassland objects in blue with the selected sample overlaid in red.

This Grafström and Tillé (2013) method thus ensures that the sample is well spread across the Swiss territory. This sampling design is particularly efficient if the variable of interest  $y_i$  is linearly dependent on the auxiliary variables contained in  $\mathbf{x}_i$ . If the sample is balanced on  $\mathbf{x}_i$ , the variance of the estimator of the total of  $\mathbf{x}_i$  is null or almost null. Generally it remains a small variance due to the ‘rounding problem’ that occurs because an exact balancing is often not possible. In a sampling design that is balanced on  $\mathbf{x}_i$ , the variance of the total estimator of the variable of interest  $y_i$  is not null. However this variance is reduced by the part of  $y_i$  that is explained by  $\mathbf{x}_i$ . Spreading also improves the efficiency when the data are autocorrelated. In spatial sampling, it is likely that two neighbouring units are often similar. To avoid collecting redundant information, neighbouring units should not be measured (Grafström et al. 2012).

### 6.3 Balanced rounding of the number units in the polygons

In the polygons, the  $n_{k,i}$  computed in Sect. 5.4 are not integer values. We therefore have to ‘round’ the  $n_{k,i}$ . The only way of satisfying exactly the inclusion probabilities is to round the  $n_{k,i}$  randomly. We will round them in such a way that the properties of balancing of the sample are preserved. This includes defining the same balancing variables  $\mathbf{x}_{k,i}$  at the level of the polygons

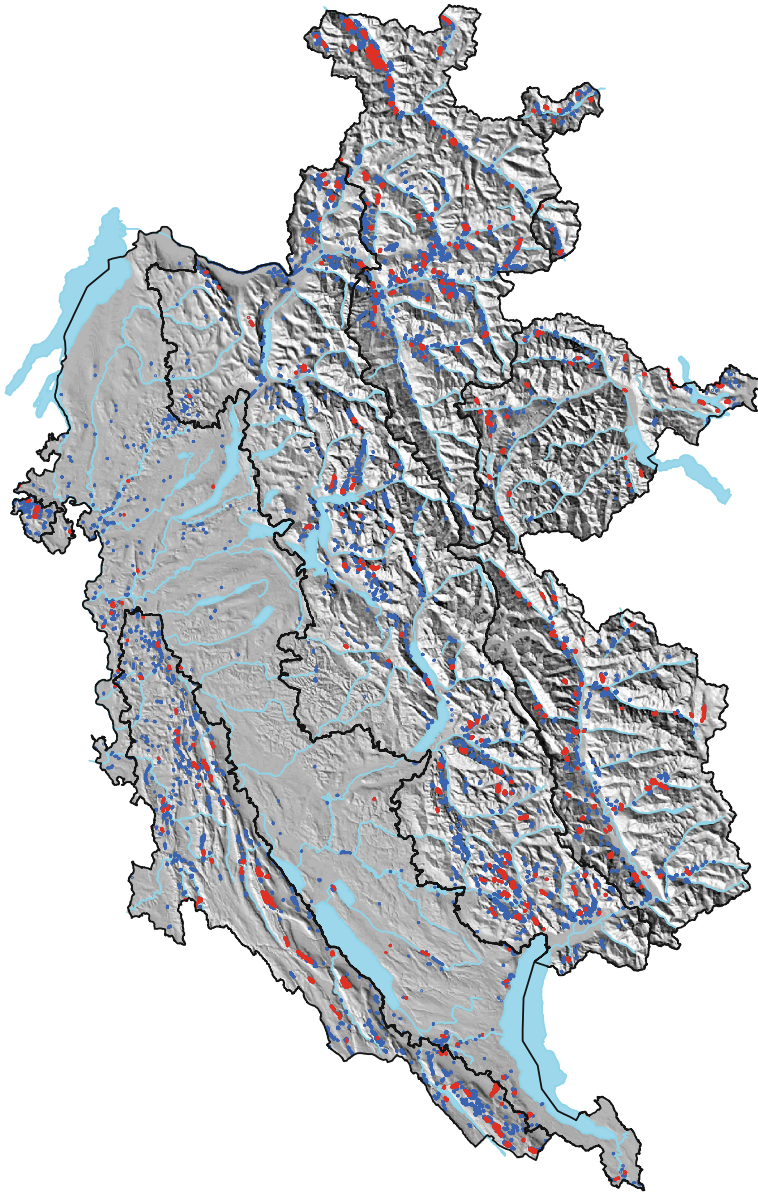
$$\mathbf{x}_{k,i} = \left( \frac{n_{k,i}}{n_i} \pi_{1,i}, A(P_k), A(P_k) \times R(P_k), D_k, D_k^2, \right. \\ \left. \text{and all the variables with the area occupied by each type of vegetation} \right), k \in O_i.$$

where variable  $D_k$  is the degree-days of the growing season in the polygon and  $D_k^2$  is the square of  $D_k$ . If  $\tilde{n}_{k,i}$  denotes the rounded value of  $n_{k,i}$  to one of its two closest integers, we can produce the balancing equations

$$\sum_{i \in S_1} \sum_{k \in O_i} \frac{\tilde{n}_{k,i}}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}} \approx \sum_{i \in S_1} \sum_{k \in O_i} \frac{n_{k,i}}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}} = \sum_{i \in S_1} \frac{\mathbf{x}_i}{\pi_{1,i}} \approx \sum_{i=1}^M \mathbf{x}_i. \tag{6}$$

We randomly round  $n_{k,i}$  in order to satisfy the inclusion probabilities. The expectation of  $\tilde{n}_{k,i}$  must thus satisfy

$$E(\tilde{n}_{k,i}) = n_{k,i}.$$



**Fig. 4** Spatial distribution of the national inventory of dry grassland objects across six biogeographical regions in Switzerland. The selected sample in *red* is superimposed on the inventory data frame in *blue*. The map is a *shaded* relief that includes the main rivers and lakes (Color figure online)

Let  $\phi_k = n_{k,i} - \lfloor n_{k,i} \rfloor$  and  $I_{k,i} = \tilde{n}_{k,i} - \lfloor n_{k,i} \rfloor$ , where  $\lfloor n_{k,i} \rfloor$  denotes the largest integer not greater than  $n_{k,i}$ . The balancing equations given in (6) can be rewritten as

$$\sum_{i \in S_1} \sum_{k \in O_i} \frac{I_{k,i} + \lfloor n_{k,i} \rfloor}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}} \approx \sum_{i \in S_1} \sum_{k \in O_i} \frac{\phi_k + \lfloor n_{k,i} \rfloor}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}}$$

or equivalently

$$\sum_{i \in S_1} \sum_{k \in O_i} \frac{I_{k,i}}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}} \approx \sum_{i \in S_1} \sum_{k \in O_i} \frac{\phi_k}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}}$$

If we pose

$$\mathbf{z}_{k,i} = \frac{\phi_k}{n_{k,i}} \frac{\mathbf{x}_{k,i}}{\pi_{1,i}},$$

we can write

$$\sum_{i \in S_1} \sum_{k \in O_i} \frac{\mathbf{z}_{k,i} I_{k,i}}{\phi_k} \approx \sum_{i \in S_1} \sum_{k \in O_i} \mathbf{z}_{k,i},$$

which is also a problem of balanced sampling that can be solved using the method of Grafström and Tillé (2013). Once  $I_{k,i}$  is selected randomly, we can compute the  $\tilde{n}_{k,i} = I_{k,i} + \lfloor n_{k,i} \rfloor$ . We have now computed the (integer) numbers of circular plots  $\tilde{n}_{k,i}$  that we have to select in each polygon. The balancing equations are then satisfied in the best possible manner.

### 6.4 Selection of the plots in the polygons

Once the  $\tilde{n}_{k,i}$  are computed, the plots can be selected in the polygons. To optimally spread the plots within a polygon, the Reversed Randomized Quadrant-Recursive Raster method is applied (Theobald et al. 2007). The method transforms a raster representation of the two-dimensional polygon into a one-dimensional list of grid cells. The grid cells are considered as plots as the resolution of the raster was 50 cm. The list is subjected to complex reorganizations before and after being systematically sampled. The result is a spatially balanced sample within the polygon. The reorganization technique also enables the user to consecutively drop plots or select new ones while maintaining optimal spreading. We used this technique to select two additional reserve plots within each object and three additional ones within each polygon.

The first category was assigned to the polygons in a way to optimally preserve the balanced character of the sample. The second category simply represents additional random plots within a polygon which maintains the spatial spreading. In the way of defining reserve plots, the sampling procedure of Theobald et al. (2007) should be very useful in future survey runs if the sampling intensity has to be flexibly adapted to changing budgets.



## 6.5 Rules for choosing a reserve plot in the field

Non target areas within the sample frame and practical problems in the field may adversely affect the efficiency of the sample and some flexibility is needed to deal with these situations, e.g. if a selected sample plot cannot be accessed because the terrain is too steep or the vegetation too dense. If plots have been recently mowed or grazed, they cannot be recorded in sufficient detail. Dense hedgerows can only be recorded from outside, while other plots may be located on land-cover types (i.e. fields, forests, houses or roads) which are not in the focus of the survey. Such plots may considerably reduce the effective sample size as either no data can be recorded or data may be missing, of low accuracy or less relevant. Choosing a plot in the nearby dry grassland, i.e. shifting it, is no solution as it increases the sampling intensity in the respective border zones and thus introduces bias. The effect is particularly pronounced if the shifting is frequently towards a particular vegetation type. Shifting in this survey is thus restricted to cases where the delineation is systematically displaced from the true position of the grassland due to problems with photogrammetry. The effect of shifting is then supposed to be neutral as it restores the expected sampling intensity at the grassland border.

For all the remaining cases, we use the reserve plots to be additionally recorded in the field. In principle, if the survey problem did not concern a particular type of vegetation a plot could be simply replaced by a reserve one. However, most survey problems in this study are supposed to be systematically linked to a particular type of vegetation. In this case simple replacement would cause bias due to the systematic exclusion of this aspect. Hence, a reserve plot is always selected in addition and the original plot, e.g. a dense hedgerow, is maintained and recorded as well as possible. The missing or impaired data may be imputed from auxiliary information (topography, bio-climate, spectral data) or by the nearest neighbour method. If a plot is not recordable at all because the terrain is too steep or the vegetation too dense, we still try to record the land-cover type and the reason why so little information is provided. The same basic recording is done for plots that do not fit the survey focus. The simple information about the land-cover type is required to estimate its proportion in the sample frame.

The selection of a reserve plot in the field is governed by the following rules: First, the surveyor chooses from the two optimally balanced reserve plots. For that, the surveyor looks for a plot in the same polygon. If no such plot exists, the plot which involves the least walking effort to be reached is chosen. If both plots are located in the same polygon the plot which was drawn first by the sampling method of [Theobald et al. \(2007\)](#) has to be selected first to retain the spatial spreading. If the two balanced reserve plots are already used, we record one of the three additional reserve plots in the current polygon. Their order of being selected is again defined by the order as they were sampled.

The purpose of the reserve plots is to maintain the effective sample size. In fact, the sample size is increased since unfavourable plots are not replaced but supplemented. Consequently, the inclusion probabilities have to be adjusted at the level of the polygon. This methodology has been developed to maintain the intended sample size. It also helps to preserve the balance between field work and travel time. If one of the sample

plots cannot be surveyed in detail, the surveyor must select and survey one of the reserve plots.

## 7 Estimation of a total

In order to estimate the total, the two stages of the sampling design, i.e. the sampling design of the objects and that within the object, must be considered. Let  $\check{n}_{k,i}$  be the final number of circular plots selected in polygon  $P_k$ . Quantity  $\check{n}_{k,i}$  includes possible nonresponse and recorded reserve plots. Quantity  $\check{n}_i$  is an integer number of circular plots in object  $O_i$ . It is the sum of the number of circular plots  $\check{n}_{k,i}$  selected in each polygon, i.e.

$$\check{n}_i = \sum_{k \in O_i} \check{n}_{k,i}.$$

Within the objects, we use several hypotheses to simplify the computation of variance:

1. The circular plots are assumed to be non-overlapping. Indeed, in practice, the probability of selecting two overlapping plots is very small.
2. Within the objects, the number of possible plots is very important, so we assume that the objects are sampled with replacement. The impact of this hypothesis is not very large since the finite population correction is very small. We consider that the sample is selected with replacement and with unequal probabilities of drawing and can thus use the Hansen–Hurwitz estimator (Hansen and Hurwitz 1949) in the objects.
3. Moreover, we consider that  $\check{n}_i$  is not random, which means that the total is computed and estimated conditionally to  $\check{n}_i$  (on conditional inference see among others Rao 1985; Valliant 1993). The number of plots (i.e sampling units) in polygon  $P_k$  (or in object  $O_i$ ) is supposed to be equal to  $A(P_k)/A(Q)$  (or  $A(O_i)/A(Q)$ ). This number is not integer but this is without consequence. It follows that the probability of selecting a circular plot in polygon  $P_k$  if  $\check{n}_i$  plots have been selected in object  $O_i$  is equal to

$$\Pr(Q_r \text{ is selected} | \check{n}_i) = \pi_{2,r|k} \frac{\check{n}_i}{n_i} = \frac{A(Q)n_{k,i}\check{n}_i}{n_i A(P_k)}.$$

Now, if we assume that the sampling design is with replacement in each object, the probability of selecting a particular plot for each drawing of a design with replacement is given by

$$p_r = \frac{\Pr(Q_r \text{ is selected} | \check{n}_i)}{\check{n}_i} = \frac{A(Q)n_{k,i}}{n_i A(P_k)},$$

for all circular plots. Since  $n_{k,i} = n_i A(P_k)R(P_k)/[A(O_i)R(O_i)]$ , we can also write

$$p_r = \frac{A(Q)}{n_i A(P_k)} \frac{n_i A(P_k)R(P_k)}{A(O_i)R(O_i)} = \frac{A(Q) R(P_k)}{A(O_i)R(O_i)}.$$

Consider now the variable of interest  $y_r$  measured in plot  $Q_r$ . Value  $y_r$  can typically be the area occupied by a particular plant in plot  $Q_r$ . The Hansen–Hurwitz estimator of the total area occupied by this plant in object  $O_i$  is given by:

$$\hat{Y}_i = \frac{1}{\check{n}_i} \sum_{k \in O_i} \sum_{r \in S_{2,k}} \frac{y_r}{p_r} = \frac{A(O_i)R(O_i)}{A(Q)\check{n}_i} \sum_{k \in O_i} \frac{1}{R(P_k)} \sum_{r \in S_{2,k}} y_r,$$

where  $S_{2,k}$  is the sample of circular plots selected in polygon  $P_k$ . It is, however, preferable to use a ratio-type estimator in the object (see [Särndal et al. 1992](#), p. 254). Notice that when  $y_r = A(Q) = 10 \text{ m}^2$  for all  $Q_r$ , then we obtain the Hansen–Hurwitz estimator of  $A(O_i)$  that is equal to

$$\hat{A}(O_i) = \frac{A(O_i)R(O_i)}{A(Q)\check{n}_i} \sum_{k \in O_i} \frac{1}{R(P_k)} \sum_{r \in S_{2,k}} A(Q) = \frac{A(O_i)R(O_i)}{\check{n}_i} \sum_{k \in O_i} \frac{\check{n}_{k,i}}{R(P_k)}.$$

A ratio estimator of  $Y_i$  is thus given by

$$\hat{Y}_{Ri} = \frac{A(O_i)}{\hat{A}(O_i)} \hat{Y}_i = \frac{A(O_i)}{A(Q)} \frac{\sum_{k \in O_i} 1/R(P_k) \sum_{r \in S_{2,k}} y_r}{\sum_{k \in O_i} \check{n}_{k,i}/R(P_k)}.$$

If  $y_r$  is the area occupied by a plant in plot  $Q_r$ , then  $\hat{Y}_{Ri}$  gives an estimate of the area occupied by this plant in object  $O_i$  provided that object  $O_i$  has been selected as primary unit.

The estimator of the total in the population is thus

$$\hat{Y} = \sum_{i \in S_1} \frac{\hat{Y}_{Ri}}{\pi_{1,i}}.$$

The weight  $w_r$  for estimator  $\hat{Y}$  of a particular plot  $Q_r$  belonging to polygon  $P_k$  of object  $O_i$  is thus given by

$$w_r = \frac{A(O_i)}{A(Q)\pi_{1,i} R(P_k) \sum_{k \in O_i} \check{n}_{k,i}/R(P_k)}.$$

The estimator can thus be written as a linear combination of the observations:

$$\hat{Y} = \sum_{k \in S} w_r y_r,$$

where  $S$  is the sample of plots. Once the weights have been computed, the total can be estimated for any variable of interest.

### 8 Variance and variance estimation

In a two-stage sampling design, an unbiased estimator of variance can be written

$$v(\widehat{Y}) = v_1 + v_2,$$

where

$$v_1 = \sum_{i \in S_1} \sum_{j \in S_1} \frac{\widehat{Y}_i \widehat{Y}_j}{\pi_{1,i} \pi_{1,j}} \frac{\pi_{1,ij} - \pi_{1,i} \pi_{1,j}}{\pi_{1,ij}},$$

$$v_2 = \sum_{i \in S_1} \frac{v(\widehat{Y}_i)}{\pi_{1,i}}$$

and  $\pi_{1,ij}$  is the probability of jointly selecting units  $i$  and  $j$ . Estimator  $v(\widehat{Y}_i)$  must be an unbiased (or approximately unbiased) estimator of the variance of  $\widehat{Y}_i$  given that object  $i$  has been selected.

Notice that  $v_1$  cannot be considered as an estimator of the variance in the first stage, and  $v_2$  cannot be considered as an estimator of the variance in the second stage. Indeed, a large part of the variance in the second stage is captured by  $v_1$ . For this reason, the value of  $v_2$  is generally very small compared to  $v_1$  (see e.g. Tillé 2001, pp. 175–177, or Särndal et al. 1992, pp. 137–139).

Assuming that the sampling design in the objects is with replacement, the variance of the Hansen–Hurwitz estimator (Hansen and Hurwitz 1949), denoted  $\widehat{Y}_i$  here, is given by

$$\text{var}(\widehat{Y}_i | S_1) = \frac{1}{\check{n}_i} \sum_{k \in O_i} \sum_{r \in P_k} p_{k(r)} \left( \frac{y_r}{p_{k(r)}} - Y_i \right)^2.$$

This variance can be estimated by

$$v(\widehat{Y}_i) = \frac{1}{\check{n}_i(\check{n}_i - 1)} \sum_{k \in O_i} \sum_{r \in S_{2,k}} \left( \frac{y_r}{p_{k(r)}} - \widehat{Y}_{Ri} \right)^2$$

$$= \frac{1}{\check{n}_i(\check{n}_i - 1)} \sum_{k \in O_i} \sum_{r \in S_{2,k}} \left( \frac{R(O_i)A(O_i)y_r}{O(Q)R(P_k)} - \widehat{Y}_{Ri} \right)^2.$$

Obviously, due to the hypothesis that the sampling design is with replacement, the variance estimator  $v(\widehat{Y}_i)$  overestimates the variance. Indeed, sampling with replacement does not take the spreading of the sampling design into account. The spreading decreases the variance but due to the small sample sizes in the objects, the effect of

spreading is probably not very significant. Moreover, the incidence of this overestimation is relatively weak since  $v_2$  is small compared to  $v_1$ .

The estimation of  $v_i$  is more intricate:

$$v_1 = \sum_{i \in S_1} \sum_{j \in S_1} \frac{\widehat{Y}_i \widehat{Y}_j}{\pi_{1,i} \pi_{1,j}} \frac{\pi_{1,ij} - \pi_{1,i} \pi_{1,j}}{\pi_{1,ij}}$$

Here, the sampling design is so complex that it is not possible to compute the joint inclusion probability for the first stage. Grafström and Tillé (2013) propose a heuristic expression to estimate the variance that takes the doubly balanced effect into account. This estimator is based on the fact that the cube method is applied locally and has been validated by a large set of simulations. The variance is then estimated locally by computing a deviation to the mean of the nearest neighbours computed with the Euclidean distance on the coordinates. Following Grafström and Tillé (2013), we thus suggest using an alternative estimator for samples that are both spread and balanced:

$$v_{1ALT} = \frac{n}{n-p} \frac{p+1}{p} \sum_{k \in S_1} (1 - \pi_{1,i}) \left( \frac{e_i}{\pi_{1,i}} - \bar{e}_i \right)^2, \tag{7}$$

where  $p$  is the number of balancing variables,

$$\bar{e}_i = \frac{\sum_{\ell \in B_i} (1 - \pi_{1\ell}) \frac{e_\ell}{\pi_{1\ell}}}{\sum_{\ell \in B_i} (1 - \pi_{1\ell})}$$

and  $B_i$  is the set of the  $p + 1$  closest objects of  $i$  in the sample of objects (including  $i$  itself). The  $\bar{e}_i$  is a local mean computed in a neighbourhood of  $i$  of the residuals

$$e_i = \widehat{Y}_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}},$$

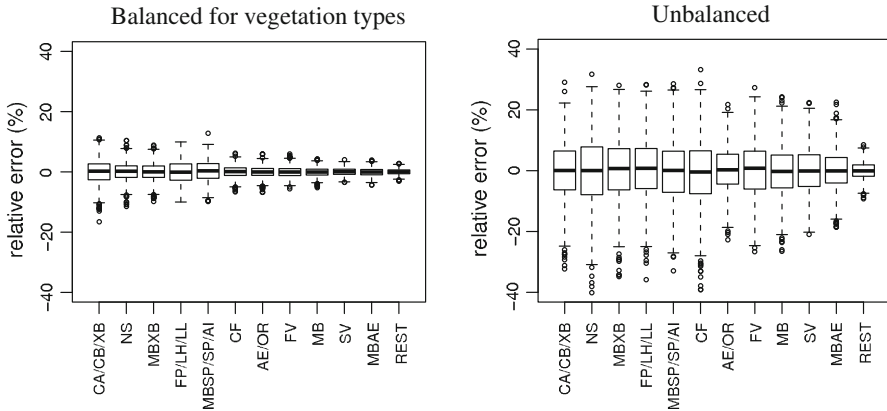
where  $\mathbf{x}_i$  is the vector of auxiliary information available for unit object  $i$ , and  $\widehat{\boldsymbol{\beta}}$  is the vector of regression coefficients estimated as follows:

$$\widehat{\boldsymbol{\beta}} = \left[ \sum_{\ell \in S} (1 - \pi_\ell) \frac{\mathbf{x}_\ell \mathbf{x}_\ell^\top}{\pi_\ell \pi_\ell} \right]^{-1} \sum_{i \in S} (1 - \pi_i) \frac{\mathbf{x}_i \widehat{Y}_i}{\pi_i \pi_i}$$

### 9 Discussion: advantages of the sampling design

The sampling design offers several advantages:

- The circular plots are selected with inclusion probabilities proportional to the rareness index.
- The design has two stages. The plots are clustered into objects.



**Fig. 5** Accuracy of proportion estimates for the 11 target vegetation types in the two sampling design. On the *left* the sampling design is balanced. On the *right* the sampling design is not balanced. The figures result from 1,000 draws of the survey design. The relative errors are calculated as follows: (real proportion-estimated proportion)  $\times$  100 / real proportion

- The methods of [Theobald et al. \(2007\)](#) and [Grafström and Tillé \(2013\)](#) ensure that the sample is well spread in geographical space.
- The method of [Grafström and Tillé \(2013\)](#) ensures that the sample is balanced, i.e. the Horvitz–Thompson estimator of the totals of the balancing variables is almost equal to the population totals.

A simulation with 1,000 draws of the proposed survey design shows much higher accuracy of the corresponding proportion estimates (Fig. 5) than applying the same simulation to a survey design without balancing for the vegetation types. The balanced sampling design reduces the interquartile ranges of the box-plots by a factor of 2.6–6.4. The mean reduction factor is 4.2.

The sampling design described here achieves all the objectives of long-term monitoring. The sampling design may seem complex, but a simple weight  $w_r$  can be computed for any plot. These weights mainly depend on the rareness index. They enable a simple extrapolation of the sample to the population or to an area of interest such as a biogeographical region. Despite the complexity of the sampling design, the variance can be estimated and confidence intervals constructed. We developed this sampling design during a long process of evaluation of what is required to monitor the national inventory of dry grasslands. We hope to have developed an observation strategy that is as efficient as possible.

**Acknowledgments** We thank Ariel Bergamini for the contributions to the content and form of the paper and Silvia Dingwall for a thorough revision of the English of the manuscript. Moreover, we thank two reviewers for several encouraging and constructive comments that enabled us to improve the quality of this paper. We are also very grateful to the Swiss Federal Office for the Environment FOEN for funding this study.

## References

- Albert CH, Yoccoz NG, Edwards TC, Graham CH, Zimmermann NE, Thuiller W (2010) Sampling in ecology and evolution—bridging the gap between theory and practice. *Ecography* 33(6):1028–1037
- Bankier M (1988) Power allocations: determining sample sizes for subnational areas. *Am Stat* 42:174–177
- Bergamini A, Holderegger R (2012) Die Wirkungskontrolle Biotopschutz Schweiz: Ein Monitoringprogramm im Aufbau. *Nat Paysage Nat Landschaft Insid* 1(12):14–20
- Bergamini A, Peintinger M, Fakheran S, Moradi H, Schmid B, Joshi J (2009) Loss of habitat specialists despite conservation management in fen remnants 1995–2006. *Perspect Plant Ecol Evol Syst* 11(1):65–79
- Chauvet G (2009) Stratified balanced sampling. *Surv Methodol* 35:115–119
- Deville J-C, Särndal C-E (1992) Calibration estimators in survey sampling. *J Am Stat Assoc* 87:376–382
- Deville J-C, Tillé Y (2004) Efficient balanced sampling: the cube method. *Biometrika* 91:893–912
- Diekmann M (2003) Species indicator values as an important tool in applied plant ecology—a review. *Basic Appl Ecol* 4(6):493–506
- Eggenberg S, Dalang T, Dipner M, Mayer C (2001) Cartography and evaluation of dry grasslands sites of national importance, vol 325. Swiss Agency for the Environment, Forests and Landscape (SAEFL), Bern, p 251
- Ellenberg H, Webwe HE, Düll R, Wirth V, Werner W, Paulissen D (1992) Zeigerwerte von Pflanzen in Mitteleuropa. *Scripta Geobotanica* 18:258
- Feldmeyer-Christe E, Ecker K, Kuchler M, Graf U, Waser L (2007) Improving predictive mapping in Swiss mire ecosystems through re-calibration of indicator values. *Appl Veg Sci* 10(2):183–192
- Graf U, Wildi O, Kuchler M, Ecker K (2010) Five-year changes in Swiss mire vegetation. *Bot Helv* 120(1):15–27
- Grafström A, Lundström NLP, Schelin L (2012) Spatially balanced sampling through the pivotal method. *Biometrics* 68(2):514–520
- Grafström A, Tillé Y (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 14(2):120–131
- Grujter Jd, Bierkens M, Brus D, Knotters M (2006) Sampling for natural resource monitoring. Springer, Berlin, Heidelberg, New York
- Hansen MH, Hurwitz WN (1949) On the determination of the optimum probabilities in sampling. *Ann Math Stat* 20:426–432
- Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 47:663–685
- Kapfer J, Grytnes J-A, Gunnarsson U, Birks HJB (2011) Fine-scale changes in vegetation composition in a boreal mire over 50 years. *J Ecol* 99:1179–1189
- Klaus G (2007) Biodiversität in Feuchtgebieten. *Hotspot* 15:3–5
- Koch MA, Scheriau C, Schupfner M, Bernhardt KG (2011) Long-term monitoring of the restoration and development of limestone grasslands in north western germany: vegetation screening and soil seed bank analysis. *Flora* 206(1):52–65
- Lachat T, Pauli D, Gonseth Y, Klaus G, Scheidegger C, Vittoz P, Walter T (2010) Wandel der Biodiversität in der Schweiz. Ist die Talsohle erreicht? Bristol Stiftung, Haupt, Zürich, p 435
- Landolt E, Bäumler B, Erhardt A, Hegg O, Klötzli F, Lämmler W, Nobis M, Rudmann-Maurer K, Schweingruber FH, Theurillat J-P, Urmi E, Vust M, Wohlgenuth, T (2010) Flora indicativa. Ökologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen. 2. völlig neu bearbeitete und erweiterte Auflage der Ökologischen Zeigerwerte zur Flora der Schweiz (1977). Haupt, Bern Stuttgart Wien
- Legendre P, Legendre L (1998) Numerical ecology, 2nd edn. Elsevier, New York
- Moser D, Gyax A, Bäumler B, Wyler N, Palese R (2002) Rote Liste der gefährdeten Farn- und Blütenpflanzen der Schweiz. BUWAL-Reihe « Vollzug Umwelt », p 118
- Odland A (2009) Interpretation of altitudinal gradients in south central norway based on vascular plants as environmental indicators. *Ecol Indic* 9(3):409–421
- Overton WS, Stehman SV (1996) Desirable design characteristics for long-term monitoring of ecological variables. *Environ Ecol Stat* 3(4):349–361
- Rao JNK (1985) Conditional inference in survey sampling. *Surv Methodol* 11:15–31
- Särndal C-E, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer, New York

- Theobald DM, Stevens DL, White D, Urquhart NS, Olsen AR, Norman JB (2007) Using GIS to generate spatially balanced random survey designs for natural resource applications. *Environ Manag* 40(1):134–146
- Thompson WL, Miller AE, Mortenson DC, Woodward A (2011) Developing effective sampling designs for monitoring natural resources in alaskan national parks: an example using simulations and vegetation data. *Biol Conserv* 144:1270–1277
- Tillé Y (2001) *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod, Paris
- Tillé Y (2006) *Sampling algorithms*. Springer, New York
- Tillé Y, Matei A (2012) *Sampling: survey sampling*. R package version 2.5, <http://CRAN.R-project.org/package=sampling>
- Valliant R (1993) Poststratification and conditional variance estimation. *J Am Stat Assoc* 88(421):89–96
- Zimmermann NE, Kienast F (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *J Veg Sci* 10(4):469–482

## Author Biographies

**Yves Tillé** is currently Professor at the University of Neuchatel in Switzerland. He holds a Ph.D. in Statistics of the University of Brussels. He was previously an assistant professor at the Ecole Nationale de la Statistique et de l'Analyse de l'Information (Rennes, France). His research interests are in survey sampling and estimation in finite populations. His main contributions deal with sampling algorithms, calibration, resampling method in finite population, and repeated sampling.

**Klaus Ecker** is a research scientist at the Swiss Federal Research Institute WSL. He holds a Ph.D. in vegetation science from the University Vienna and is active in long-term monitoring programs on habitat conservation in Switzerland. His research focuses on sampling issues, vegetation change and remote sensing based ecological modeling.