

ELECTRONIC SUPPLEMENTARY MATERIAL

ESM1- Supplementary Material

ARTICLE TITLE

Monitoring habitat conservation status and assessing biodiversity: The Dehesa case

JOURNAL

Environmental Monitoring and Assessment

AUTHOR

Luis Ambrosio Flores

AFFILIATION

Universidad Politécnica de Madrid, Madrid, Spain

CORRESPONDING AUTHOR

Luis Ambrosio Flores

E-MAIL

luis.ambrosio@upm.es

1. Introduction

This supplementary material provides detailed descriptions of the statistical methods proposed in the main text for monitoring conservation status, as well as carbon removal and biodiversity. Tree, grassland, and shrub components are each considered distinct statistical populations, defined as sets of elements that share a common definition (e.g., tree, grass, or shrub), but whose individual characteristics vary across the habitat. Each indicator of conservation status for these populations is treated as a survey variable defined over spatial sampling units and observed at successive monitoring occasions. Habitat-level indicators are expressed as aggregates (totals and ratios) of these variables and are estimated from probabilistic samples of spatial units within a design-based inference framework.

A key element of this framework is the design of a probability sampling scheme to select a random sample of spatial units where data on the survey variables are collected. This scheme assigns each sampling unit a known probability of inclusion in the sample. The estimator is defined as a function of the observed data, and inference is based on its sampling distribution induced by the sampling design. Design-based estimators that rely solely on sampled ground data are unbiased; that is, their sampling distribution is centered on the true population value of the aggregate being estimated. However, they are often inefficient in the sense that the variance of their sampling distribution can be high when the sample size is small.

To improve efficiency, we complement ground data on the survey variable with information on auxiliary variables, such as Earth observation (EO) data, within a model-assisted framework. In this framework, a statistical model is specified to relate the survey variable to the auxiliary variables and is used to predict survey-variable values for sampling units not included in the sample. The model-assisted estimator combines the aggregate of these predictions over the population with a design-weighted correction based on the prediction residuals observed in the sample. Although it may be biased in small samples, it is design-consistent, meaning that its sampling distribution becomes increasingly concentrated around the true finite-population indicator value as the sample size increases.

However, inference remains design-based because it relies on the sampling distribution induced by the probability-sampling design, applied to the model residuals. As a result, model-assisted estimators are robust in the sense that their validity does not depend on the correctness of the model: if the model is correctly specified, the residuals are small, the variance of the sampling distribution is reduced, and the gain in efficiency relative to estimators based solely on ground data is substantial; otherwise, the efficiency gain is limited.

We focus on model-assisted estimation. For continuous survey variables, we use linear models. For categorical survey variables, we use logistic regression models: binary logistic models when the variable is dichotomous, and multinomial logistic models when it is polytomous. In all cases, the estimators of the model parameters are design-

consistent; as a result, the corresponding model-assisted indicator estimators are also robust and design-consistent.

2. Probability sampling design

We adopt a multistage sampling design because it offers flexibility to adapt both the number of stages and the probabilistic scheme to the specific characteristics of the populations under study (trees, grassland, shrubs, and fauna). The choice of estimators depends on data availability: when EO data are available alongside sampled ground data, we employ design-based model-assisted estimators. In cases where only sampled ground data are available, we use design-based estimators.

To support ground data collection, both list frames and area frames are used. In the first stage, the sampling frame consists of a list (e.g., a census or an administrative register) of the M farms (Dehesas) that make up the habitat. A sample of m farms is selected through m independent draws with replacement and with probabilities proportional to size (PPS). For farm i , with area A_i , let A_i/A denote its selection probability on any one draw, where $A = \sum_{i=1}^M A_i$ is the total area of the habitat.

The estimation procedures used below are based on the Hansen–Hurwitz formulation and therefore use the draw-level selection probabilities, A_i/A . Sampling with replacement is adopted because it simplifies variance estimation, which is particularly advantageous in multistage designs. The resulting loss of precision is negligible when, as in our case, the number M of farms in the population is large relative to the first-stage sample size m .

In the second stage, an area sampling frame based on Sentinel-2 satellite imagery and LiDAR data is used. Within each selected farm $i = 1, 2, \dots, m$, pixels of $10 \text{ m} \times 10 \text{ m}$ are stratified into H strata using tree density derived from LiDAR as the stratification variable. Let N_{ih} denote the number of pixels in stratum $h = 1, 2, \dots, H$ of the farm i .

Within each stratum h of each selected farm i , a sample of n_{ih} pixels is selected without replacement, with inclusion probability $\pi_{p|ih}$ for pixel $p = 1, 2, \dots, N_{ih}$ (conditional on the selection of farm i). For simple random sampling, $\pi_{p|ih} = n_{ih}/N_{ih}$.

This two-stage sample is sufficient to estimate most indicators of the arboreal layer. For the remaining monitored layers, in addition to these first and second stages common to all layers, a third-stage sample is required for cost-efficiency reasons. Within each pixel $p = 1, 2, \dots, n_{ih}$ selected at the second stage, a sample of n_{ihp} quadrats is drawn without replacement from the N_{ihp} possible quadrats contained within the pixel. The inclusion

probability of a quadrat in the sample of the selected farm i is $\pi_{pq|ih} = \pi_{p|ih} \pi_{q|ihp}$. For simple random sampling $\pi_{q|ihp} = n_{ihp} / N_{ihp}$.

Line-intersect Sampling (LIS)

For certain faunal groups, particularly birds and other highly mobile vertebrates, area-based sampling may be difficult to apply because individuals move across spatial sampling units and population boundaries are often difficult to define. In such cases, line transects and circular variable plots have been proposed in the literature (Thompson, 2012). Here, we consider a two-stage sampling design, with the sample of farms common to all monitored layers at the first stage. At the second stage, a set of one or more transects is selected within each sampled farm, with independent transect sets selected at successive monitoring occasions.

Let L_{it} denote the total length of the set of non-overlapping transects selected in sampled farm i at monitoring occasion t . We assume that animal j is detected when the observer passes within a distance r_{ijt} of its location. Applying the endpoint convention described by de Vries (1986, pp. 247–248), the set of transect placements that results in the inclusion of animal j has area $2r_{ijt}L_{it}$, provided that the inclusion regions associated with the transects do not overlap and that $2r_{ijt}L_{it} \leq A_i$. Consequently, conditional on the selection of farm i , the second-stage inclusion probability of animal j at occasion t is $\pi_{j|it} = 2r_{ijt}L_{it} / A_i$. In Numerical Example 6, two non-overlapping transects of equal length are used in each sampled farm at each monitoring occasion, and the two-transect sets are selected independently between occasions.

Panel sampling

For change estimation, the same first-stage units (farms) are revisited at each monitoring occasion within a fixed monitoring cycle. Area-based indicators follow a strict panel design at the subsequent sampling stages: the same second- and third-stage units (e.g., pixels, trees or quadrats, depending on the indicator) are revisited. For LIS-based indicators, however, a new set of transects is selected at each monitoring occasion, resulting in independent occasion-specific second-stage LIS samples.

3. Estimators

To estimate habitat-level indicators of either status conservation, productivity, functional traits, or fauna biodiversity, we consider both design-based and model-assisted estimators. The former relies solely on sampled data for the survey variable associated with the indicator, whereas the latter additionally incorporates Earth observation (EO) data.

3.1 Design-based estimation

To estimate the required indicators in either trees, grassland, shrub, or faunal populations, we use a same class of estimators where survey variable values observed

in the sampling units are weighted by the inverse of the inclusion probability and aggregated at the sampling stage level. We use a two-stage area sampling for trees, a three-stage area sampling for grassland and shrub, and a two-stage LIS sampling for fauna diversity.

3.1.1 Totals

Let Y_U denote the population total of a survey variable for population U , such as trees, grassland, shrub, or fauna. Under first-stage PPS sampling with replacement, the

Hansen–Hurwitz estimator of Y_U is $\hat{Y}_U = A1/m \sum_{i=1}^m \hat{Y}_{Ui}/A_i$, where $\hat{Y}_{Ui}$ estimates the survey-variable total in the selected farm $i = 1, 2, \dots, m$. The sampling variance of $\hat{Y}_U$ is estimated by $\hat{V}\hat{Y}_U = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{Ui}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{Ui}/A_i \right)^2$.

For large samples, the standardized estimator converges in distribution to a normal $(\hat{Y}_U - Y_U)/\sqrt{\hat{V}\hat{Y}_U} \rightarrow N(0,1)$ (Fuller, 2009, p. 42). Accordingly, an approximate 95% confidence interval for Y_U is derived as a measure of uncertainty: $\hat{Y}_U \pm 1.96\sqrt{\hat{V}\hat{Y}_U}$.

The estimators for each ecological component are particular cases of $\hat{Y}_U$. For trees, we consider two totals: the total number of trees, Y_N , and the total number of young trees,

Y_{N_y} . An unbiased estimator of Y_N is $\hat{Y}_N = A1/m \sum_{i=1}^m \hat{Y}_{Ni}/A_i$, where $\hat{Y}_{Ni} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} y_{ihp}/\pi_{hp|i}$

with y_{ihp} denoting the number of trees observed in sampled pixel $p = 1, 2, \dots, n_{ih}$ within the stratum $h = 1, 2, \dots, H$ of sampled farm $i = 1, 2, \dots, m$. Its sampling variance is

estimated by $\hat{V}\hat{Y}_N = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{Ni}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{Ni}/A_i \right)^2$. Similarly,

$\hat{Y}_{N_y} = A1/m \sum_{i=1}^m \hat{Y}_{N_y i}/A_i$ is unbiased for Y_{N_y} , with $\hat{Y}_{N_y i} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} y_{ihpj}/\pi_{hp|i}$ and $y_{ihpj} = 1$

if tree j in pixel p of stratum h in farm i is young, and $y_{ihpj} = 0$ otherwise. Its

sampling variance is estimated by $\hat{V}\hat{Y}_{N_y} = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{N_y i}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{N_y i}/A_i \right)^2$.

For grassland and shrub, we consider three totals: total dry matter (DM), Y_{DM} , total crude protein (CP), Y_{CP} , and total shrub-covered area, Y_{SH} . An unbiased estimator of Y_l

, where l denotes DM, CP, or shrub-covered area, is $\hat{Y}_l = A1/m \sum_{i=1}^m \hat{Y}_{li}/A_i$ with

$\hat{Y}_{li} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{hp}} y_{ihpq} / \pi_{hpq|i}$. Here, y_{ihpq} denotes the amount of DM, CP, or the shrub-covered are, observed in sampled quadrat q within sampled pixel $p = 1, 2, \dots, n_{ih}$ in stratum $h = 1, 2, \dots, H$ of selected farm $i = 1, 2, \dots, m$; and $\pi_{hpq|i}$ is its inclusion probability. Its sampling variance is estimated by

$$\hat{V}\hat{Y}_l = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{li}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{li}/A_i \right)^2.$$

For fauna estimation using LIS, we consider two totals: the total number of birds, Y_B , and the total number of birds belonging to species k , Y_{B_k} . The monitoring-occasion subscript t is suppressed throughout this section. An unbiased estimator of Y_B is

$\hat{Y}_B = A1/m \sum_{i=1}^m \hat{Y}_{Bi}/A_i$ with $\hat{Y}_{Bi} = \sum_{j=1}^{n_i} 1/\pi_{j|i} = A_i/2L_i \sum_{j=1}^{n_i} 1/r_{ij}$ where n_i is the number of birds detected using the set of transects of total length L_i in farm i , and $\pi_{j|i} = 2r_{ij}L_i/A_i$ is the conditional inclusion probability of bird j . Its sampling variance is estimated by

$$\hat{V}\hat{Y}_B = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{Bi}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{Bi}/A_i \right)^2.$$

An unbiased estimator of Y_{B_k} is $\hat{Y}_{B_k} = A1/m \sum_{i=1}^m \hat{Y}_{B_{ki}}/A_i$ with

$\hat{Y}_{B_{ki}} = \sum_{j=1}^{n_i} y_{ijk}/\pi_{j|i} = A_i/2L_i \sum_{j=1}^{n_i} y_{ijk}/r_{ij}$, where $y_{ijk} = 1$ if bird j belongs to species k and $y_{ijk} = 0$ otherwise. Its sampling variance estimator is estimated by

$$\hat{V}\hat{Y}_{B_k} = A^2 1/m(m-1) \sum_{i=1}^m \left(\hat{Y}_{B_{ki}}/A_i - 1/m \sum_{i=1}^m \hat{Y}_{B_{ki}}/A_i \right)^2.$$

3.1.2 Ratios or proportions

We consider three ratio estimators: the proportion of young trees, the grassland CP-to-DM ratio, and the proportion of birds belonging to species k . These are particular cases of the general ratio estimator $\hat{R}_u = \hat{Y}_u/\hat{Y}_v$. For trees, the estimator of the proportion of young trees is $\hat{R}_{N_y} = \hat{Y}_{N_y}/\hat{Y}_N$. For grassland, the proportion of CP to DM estimator is $\hat{R}_{CP} = \hat{Y}_{CP}/\hat{Y}_{DM}$. For fauna, the estimator of the proportion of birds belonging to species k is $\hat{R}_{B_k} = \hat{Y}_{B_k}/\hat{Y}_B$.

Using a first-order Taylor linearization (Wolter, 2007; Särndal et al., 1992, Section 5.5) we obtain $\hat{R}_u \doteq R_u + a_u \hat{Y}_u + a_v \hat{Y}_v$, where $a_u = \partial \hat{R}_u / \partial \hat{Y}_u \Big|_{\hat{Y}_u=Y_u, \hat{Y}_v=Y_v} = 1/Y_v$, and

$a_v = \partial \hat{R}_u / \partial \hat{Y}_v \Big|_{\hat{Y}_u=Y_u, \hat{Y}_v=Y_v} = -Y_u/Y_v^2 = -R_u/Y_v$. Therefore, $\hat{R}_u \doteq R_u + 1/Y_v (\hat{Y}_u - R_u \hat{Y}_v)$ and

$\hat{R}_u$ is approximately unbiased for R_u , that is, $E\hat{R}_u \doteq R_u$. Its approximate sampling variance is estimated by $\hat{V}\hat{R}_u = 1/\hat{Y}_v^2 \left[\hat{R}_u^2 \hat{V}(\hat{Y}_v) - 2\hat{R}_u \hat{Cov}(\hat{Y}_u, \hat{Y}_v) + \hat{V}(\hat{Y}_u) \right]$. Here,

$$\hat{Y}_u = A1/m \sum_{i=1}^m \hat{Y}_{ui}/A_i \text{ where } \hat{Y}_{ui} = \sum_{h=1}^H \sum_{p=1}^{n_{hp}} \sum_{j=1}^{y_{ihp}} y_{ihpj} / \pi_{hp|i} \text{ for } u \equiv N_y,$$

$$\hat{Y}_{ui} = \sum_{h=1}^H \sum_{p=1}^{n_{hp}} \sum_{q=1}^{n_{hpq}} y_{ihpq} / \pi_{hpq|i} \text{ for } u \equiv CP, \text{ and } \hat{Y}_{ui} = \sum_{j=1}^{n_i} y_{ijk} / \pi_{j|i} = A_i / 2L_i \sum_{j=1}^{n_i} y_{ijk} / r_{ij} \text{ for}$$

$u = B_k$. In the LIS expression, L_i , r_{ij} , and $\pi_{j|i}$ denote the occasion-specific quantities with the monitoring-occasion subscript suppressed.

For all three cases, the sampling variance and covariance estimators are, respectively,

$$\hat{V}\hat{Y}_u = 1/m(m-1) \sum_{i=1}^m \left(A\hat{Y}_{ui}/A_i - \hat{Y}_u \right)^2 \text{ and}$$

$$\hat{Cov}(\hat{Y}_u, \hat{Y}_v) = 1/m(m-1) \sum_{i=1}^m \left(A\hat{Y}_{ui}/A_i - \hat{Y}_u \right) \left(A\hat{Y}_{vi}/A_i - \hat{Y}_v \right).$$

An approximate 95% confidence interval for R_u is $\hat{R}_u \pm 1.96\sqrt{\hat{V}\hat{R}_u}$.

3.1.3. Change estimation

The change in a habitat-level indicator over the monitoring period is estimated using a strict panel sample for area-based indicators. For LIS-based indicators, the same first-stage sample of farms is retained, whereas independent sets of transects are selected at the second stage at each monitoring occasion. Therefore, the first-stage covariance between occasion-specific estimates is retained in the habitat-level variance estimator, whereas the corresponding second-stage covariance is zero. At the farm level, conditional on the selected farm, the variance of LIS-based change is consequently estimated as the sum of the occasion-specific variances.

3.1.3.1 Totals

The change in the habitat-level total between the beginning and the end of the monitoring period is $\Delta Y_U = Y_{UT} - Y_{U1}$ and is estimated by $\Delta \hat{Y}_U = \hat{Y}_{UT} - \hat{Y}_{U1}$. Its sampling variance is estimated by $\hat{V}\Delta \hat{Y}_U = \hat{V}(\hat{Y}_{UT} - \hat{Y}_{U1}) = \hat{V}\hat{Y}_{U1} + \hat{V}\hat{Y}_{UT} - 2\hat{Cov}(\hat{Y}_{U1}, \hat{Y}_{UT})$, where $\hat{Y}_{U_t}$ and $\hat{V}\hat{Y}_{U_t}$ are defined in section 3.1.1 for $t=1, T$, and

$$\hat{Cov}(\hat{Y}_{U1}, \hat{Y}_{UT}) = 1/m(m-1) \sum_{i=1}^m \left(A\hat{Y}_{U1i}/A_i - \hat{Y}_{U1} \right) \left(A\hat{Y}_{UIT}/A_i - \hat{Y}_{UT} \right).$$

For LIS-based total estimators at the habitat level, this covariance estimator retains only the first-stage covariance induced by the repeated sample of farms; the second-stage covariance is zero because the occasion-specific transect sets are selected independently. Accordingly, conditional on the selection of a given farm, the covariance between its two occasion-specific LIS total estimators is zero. An approximate 95% confidence interval for ΔY_U is $\Delta \hat{Y}_U \pm 1.96\sqrt{\hat{V}\Delta \hat{Y}_U}$.

3.1.3.2 Ratios and proportions

Changes in the proportion of young trees, the grassland CP-to-DM ratio, and the proportion of birds belonging to species k are estimated using the general ratio-change estimator $\Delta\hat{R}_u = \hat{R}_{uT} - \hat{R}_{u1}$. This estimator is approximately unbiased and its sampling variance is estimated by $\hat{V}\Delta\hat{R}_u = \hat{V}\left(\hat{R}_{uT} - \hat{R}_{u1}\right) = \hat{V}\hat{R}_{u1} + \hat{V}\hat{R}_{uT} - 2\hat{Cov}\left(\hat{R}_{u1}, \hat{R}_{uT}\right)$.

For $t = 1, T$ the occasion-specific estimators and their sampling variance estimators are defined in Section 3.1.2. The covariance between the two ratio estimators is estimated by $\hat{Cov}\left(\hat{R}_{u1}, \hat{R}_{uT}\right) = 1/\hat{Y}_{v1}\hat{Y}_{vT} \hat{Cov}\left(\hat{\varepsilon}_{u1}, \hat{\varepsilon}_{uT}\right)$, with

$$\hat{Cov}\left(\hat{\varepsilon}_{u1}, \hat{\varepsilon}_{uT}\right) = 1/m(m-1) \sum_{i=1}^m \left(A\hat{\varepsilon}_{ui1}/A_i - \hat{\varepsilon}_{u1}\right)\left(A\hat{\varepsilon}_{uiT}/A_i - \hat{\varepsilon}_{uT}\right), \text{ where } \hat{\varepsilon}_{uit} = \hat{Y}_{u,t} - \hat{R}_{ut}\hat{Y}_{v,t}$$

$$\text{for } t = 1, T \text{ and } \hat{\varepsilon}_{ut} = A1/m \sum_{i=1}^m \hat{\varepsilon}_{uit}/A_i.$$

For trees $u \equiv N_y$, $v \equiv N$, and $\hat{\varepsilon}_{N_yit} = \hat{Y}_{N_yit} - \hat{R}_{N_yt}\hat{Y}_{N_it}$; for grassland $u \equiv CP$, $v \equiv DM$, and $\hat{\varepsilon}_{CPit} = \hat{Y}_{CPit} - \hat{R}_{CPl}\hat{Y}_{DMit}$; and for birds, $u \equiv B_k$, $v \equiv B$, and $\hat{\varepsilon}_{Bkit} = \hat{Y}_{Bkit} - \hat{R}_{Bkt}\hat{Y}_{Bit}$. For LIS-based faunal proportions at the habitat level, the covariance estimator retains only the first-stage covariance induced by the repeated sample of farms; the second-stage covariance is zero because the occasion-specific transect sets are selected independently. An approximate 95% confidence interval for $\Delta\hat{R}_u$ is

$$\Delta\hat{R}_u \pm 1.96\sqrt{\hat{V}\Delta\hat{R}_u}.$$

3.2 Model-assisted estimators

EO information includes vegetation indices derived from multispectral satellite imagery and LiDAR-derived variables such as tree height, crown structure, and shrub cover. These data are incorporated into model-assisted estimators through statistical models relating the survey variables of interest to auxiliary EO information. The models serve exclusively as auxiliary predictive tools, whereas inference remains design-based and relies on the underlying probability-sampling design. Under the usual regularity conditions, design consistency is therefore preserved even if the working model is misspecified, although any gain in precision depends on the strength of the relationship between the EO auxiliary variables and the field observations (Särndal et al., 1992).

3.2.1 Linear models for grassland and shrub estimation

For continuous survey variables, such as grassland DM or CP and shrub-covered area, we consider a linear model relating the ground measurement observed in quadrat q , within pixel p of stratum h in farm i , denoted by y_{ihpq} , to a vector of auxiliary EO variables, $\mathbf{x}_{ihpq}$. For grassland, these auxiliary variables include vegetation indices derived from remote sensing at the pixel level, so that $\mathbf{x}_{ihpq} = \mathbf{x}_{ihp}$ for every quadrat q within pixel p . For shrub, they include LiDAR-derived shrub occupancy in the parent

1×1 -m cell containing quadrat q , and may therefore vary among quadrats within the same pixel.

The model is $y_{ihpq} = \mathbf{x}_{ihpq} \boldsymbol{\beta} + \varepsilon_{ihpq}$, where $\mathbf{x}_{ihpq} = \text{row}_{1 \leq l \leq L}(x_{lihpq})$ is a row vector of length L .

Its first component is equal to 1 and represents the intercept, whereas the remaining $L - 1$ components contain the EO auxiliary values associated with quadrat q , either at the pixel level for grassland or at the parent-cell level for shrub.

The EO covariates are available for the population sampling units even when the corresponding response values y_{ihpq} are not observed. For pixels not included in the sample, only $\mathbf{x}_{ihpq}$ EO values are available. The conditional expectation of the response is modeled as $E y_{ihpq} = \mathbf{x}_{ihpq} \boldsymbol{\beta}$, and the residual, $\varepsilon_{ihpq} = y_{ihpq} - \mathbf{x}_{ihpq} \boldsymbol{\beta}$ represents the part of y_{ihpq} not explained by the EO covariates. In the model-assisted estimator, the contribution of this unexplained variation to uncertainty is captured by the design-based sampling variance estimator described below.

3.2.1.1 Totals

Model-assisted estimation of the grassland and shrub totals—total dry matter, Y_{DM} , total crude protein, Y_{CP} , and total shrub-covered area, Y_{SH} —is based on the linear model introduced above. Let Y_l , where l denotes either DM, CP, or shrub-covered (SH) area,

be the corresponding habitat-level total $Y_l = \sum_{i=1}^M \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{q=1}^{N_{ihp}} y_{ihpq}$. Because the model

includes an intercept ($x_{lih p} = 1$), the finite-population least-squares normal equations

imply that Y_l can be expressed as $Y_l = \mathbf{x} \mathbf{B}$, where $\mathbf{x} = \sum_{i=1}^M \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{q=1}^{N_{ihp}} \mathbf{x}_{ihpq}$. The finite-

population regression coefficient vector $\mathbf{B} = \text{col}_{1 \leq l \leq L}(B_l)$ is

$$\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \left(\sum_{i=1}^M \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{q=1}^{N_{ihp}} \mathbf{x}_{ihpq}^T \mathbf{x}_{ihpq} \right)^{-1} \left(\sum_{i=1}^M \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{q=1}^{N_{ihp}} \mathbf{x}_{ihpq}^T y_{ihpq} \right), \text{ with}$$

$$\mathbf{X} = \text{col}_{1 \leq i \leq M} \text{col}_{1 \leq h \leq H} \text{col}_{1 \leq p \leq N_{ih}} \text{col}_{1 \leq q \leq N_{ihp}} \text{row}(x_{lih p}) \text{ and } \mathbf{y} = \text{col}_{1 \leq i \leq M} \text{col}_{1 \leq h \leq H} \text{col}_{1 \leq p \leq N_{ih}} \text{col}_{1 \leq q \leq N_{ihp}} (y_{ihpq}) \text{ (Fuller, 2009).}$$

In principle, the auxiliary-total vector $\mathbf{x}$ and the population design matrix $\mathbf{X}$ can be computed from wall-to-wall EO information: at the pixel level for grassland and at the corresponding parent-cell level for shrub. However, the population response vector $\mathbf{y}$ is unknown because y_{ihpq} is observed only for the sampled quadrats. Consequently, the finite-population regression coefficient vector $\mathbf{B}$ is also unknown. We estimate it from the observed y_{ihpq} values and the corresponding EO covariates using the weighted design-consistent estimator

$$\hat{\mathbf{B}} = \left(A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} \mathbf{x}_{ihpq}^T \mathbf{x}_{ihpq} / \pi_{hpq|i} \right)^{-1} A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} \mathbf{x}_{ihpq}^T y_{ihpq} / \pi_{hpq|i}$$

(Fuller, 2009).

Although the design-consistent estimator $\hat{Y}_i = \mathbf{x}_i \hat{\mathbf{B}}$, could be used to estimate the habitat-level total, Y_i , computing the required auxiliary totals over all M farms may be costly and is not necessary under the adopted sampling design. Instead, auxiliary-data processing is restricted to the m farms selected at the first stage. For each selected farm

$i = 1, 2, \dots, m$ let $\mathbf{x}_i = \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{q=1}^{N_{ihp}} \mathbf{x}_{ihpq}$ denote the population total of the auxiliary variables over all pixels and quadrats in that farm. The corresponding farm-level total is estimated

using the model-assisted regression estimator $\hat{Y}_{li} = \mathbf{x}_i \hat{\mathbf{B}} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} (y_{ihpq} - \mathbf{x}_{ihpq} \hat{\mathbf{B}}) / \pi_{hpq|i} \cdot A$

first-order Taylor linearization of $\hat{Y}_{li}$ about the finite-population coefficient vector $\mathbf{B}$

yields the linearized estimator $\hat{Y}_{li} \doteq \hat{Y}_{li0} = \mathbf{x}_i \mathbf{B} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} (y_{ihpq} - \mathbf{x}_{ihpq} \mathbf{B}) / \pi_{hpq|i}$ (Särndal et

al., 1992, Section 6.6). Because the Horvitz–Thompson residual correction is design-

unbiased for the farm-level residual total, $E \hat{Y}_{li} \doteq \mathbf{x}_i \mathbf{B} + Y_{li} - \mathbf{x}_i \mathbf{B} = Y_{li}$. Thus $\hat{Y}_{li}$ is

approximately design-unbiased for Y_{li} . Its approximate design-based variance is

$$V \hat{Y}_{li} \doteq \sum_{h=1}^H V \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} (y_{ihpq} - \mathbf{x}_{ihpq} \mathbf{B}) / \pi_{hpq|i} \cdot \text{Let } \hat{\varepsilon}_{lih p} = \sum_{q=1}^{n_{ihp}} \hat{\varepsilon}_{lih p q} / \pi_{q|i h p}, \text{ where}$$

$$\hat{\varepsilon}_{lih p q} = y_{lih p q} - \mathbf{x}_{lih p q} \hat{\mathbf{B}}. \text{ The estimate of the variance is } \hat{V} \hat{Y}_{li} \doteq \sum_{h=1}^H \hat{V} \sum_{p=1}^{n_{ih}} \sum_{q=1}^{n_{ihp}} \hat{\varepsilon}_{lih p q} / \pi_{hpq|i},$$

where,

$$\hat{V} \hat{Y}_{li} \doteq \sum_{h=1}^H \left[\frac{1}{2} \sum_{p=1}^{n_{ih}} \sum_{p'=1}^{n_{ih}} \left(\left(\pi_{hp|i} \pi_{hp'|i} - \pi_{hpp'|i} \right) / \pi_{hpp'|i} \right) \left(\hat{\varepsilon}_{lih p} / \pi_{hp|i} - \hat{\varepsilon}_{lih p'} / \pi_{hp'|i} \right)^2 + \sum_{p=1}^{n_{ih}} \hat{\Lambda}_{lih p} / \pi_{hp|i} \right] \text{ with}$$

$$\hat{\Lambda}_{lih p} = \frac{1}{2} \sum_{q=1}^{n_{ihp}} \sum_{q'=1}^{n_{ihp}} \left(\left(\pi_{q|i h p} \pi_{q'|i h p} - \pi_{qq'|i h p} \right) / \pi_{qq'|i h p} \right) \left(\hat{\varepsilon}_{lih p q} / \pi_{q|i h p} - \hat{\varepsilon}_{lih p q'} / \pi_{q'|i h p} \right)^2.$$

These results are relevant to monitoring, reporting, and verification (MRV) systems requiring statistically sound estimation of farm-level environmental indicators, including carbon-related variables. The statistical properties of the model-assisted estimators considered here rely on large-sample approximations; consequently, their adequacy must be evaluated in relation to the effective sample sizes and sampling designs used in each application. When the normal approximation is adequate, uncertainty in the farm-level total can be summarized by the approximate 95%

confidence interval, $\hat{Y}_{li} \pm 1.96 \sqrt{\hat{V} \hat{Y}_{li}}$.

At the habitat level, the total is estimated by $\hat{Y}_l = A_l/m \sum_{i=1}^m \hat{Y}_{li}/A_i$. Its sampling variance is estimated by $\hat{V}(\hat{Y}_l) = 1/m(m-1) \sum_{i=1}^m (A \hat{Y}_{li}/A_i - \hat{Y}_l)^2$. For large samples, the standardized estimator, $(\hat{Y}_l - Y_l)/\sqrt{\hat{V}\hat{Y}_l} \rightarrow N(0,1)$, converges in distribution to a normal.

Accordingly, an approximate 95% confidence interval for Y_l is $\hat{Y}_l \pm 1.96\sqrt{\hat{V}\hat{Y}_l}$.

3.2.1.2 Ratios or proportions

At the habitat level, ratios are estimated as the quotient of two model-assisted estimators of population totals. For example, the grassland CP-to-DM ratio is estimated by $\hat{R}_{CP} = \hat{Y}_{CP}/\hat{Y}_{DM}$, where $\hat{Y}_{CP}$ and $\hat{Y}_{DM}$ are model-assisted totals of total CP and total DM, respectively. The grassland CP-to-DM ratio estimate is linearized as

$\hat{R}_{CP} \doteq R_{CP} + 1/Y_{DM} (\hat{Y}_{CP} - R_{CP}\hat{Y}_{DM})$. This shows that $\hat{R}_{CP}$ is approximately unbiased for R_{CP} , that is, $E\hat{R}_{CP} \doteq R_{CP}$, with approximate variance estimator

$\hat{V}\hat{R}_{CP} = 1/\hat{Y}_{DM}^2 [\hat{V}(\hat{Y}_{CP}) - 2\hat{R}_{CP}\hat{Cov}(\hat{Y}_{DM}, \hat{Y}_{CP}) + \hat{R}_{CP}^2\hat{V}(\hat{Y}_{DM})]$. Here, the two variance estimators are particular cases of $\hat{V}(\hat{Y}_l)$ given in 3.2.1.1 with l denoting either DM or

CP, and $\hat{Cov}(\hat{Y}_{DM}, \hat{Y}_{CP}) = 1/m(m-1) \sum_{i=1}^m (A \hat{\epsilon}_{CPi}/A_i - \hat{\epsilon}_{CP})(A \hat{\epsilon}_{DMi}/A_i - \hat{\epsilon}_{DM})$ is the

corresponding variance estimator. For large samples, the standardized estimator $(\hat{R}_{CP} - R_{CP})/\sqrt{\hat{V}\hat{R}_{CP}} \rightarrow N(0,1)$ converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for R_{CP} is

$$\hat{R}_{CP} \pm 1.96\sqrt{\hat{V}\hat{R}_{CP}}.$$

3.2.1.3 Change estimation

Change is estimated as described in Section 3.1.3, replacing the design-based estimators with the corresponding model-assisted estimators.

3.2.1.3.1 Totals

At the farm level, the change between monitoring occasions $t=1$ and $t=T$ is estimated by $\Delta\hat{Y}_{li} = \hat{Y}_{liT} - \hat{Y}_{li1}$, where $\hat{Y}_{li}$ is defined in section 3.2.1.1 for $t=1, T$. Its sampling

variance is estimated by $\hat{V}\Delta\hat{Y}_{li} = \hat{V}(\hat{Y}_{liT} - \hat{Y}_{li1}) = \hat{V}\hat{Y}_{li1} + \hat{V}\hat{Y}_{liT} - 2\hat{Cov}(\hat{Y}_{li1}, \hat{Y}_{liT})$, where

$\hat{V}\hat{Y}_{li}$ is defined in 3.2.1.1 for $t=1, T$, and $\hat{Cov}(\hat{Y}_{li1}, \hat{Y}_{liT}) \doteq \sum_{h=1}^H \hat{Cov}(\hat{Y}_{lih1}, \hat{Y}_{lihT})$ estimates the

covariance between the two occasion-specific farm-level estimators, with

$$\hat{Cov}(\hat{Y}_{lih1}, \hat{Y}_{lihT}) = \frac{1}{2} \sum_{p=1}^{n_h} \sum_{p'=1}^{n_h} \left(\left(\pi_{hp|i} \pi_{hp'|i} - \pi_{hpp|i} \right) / \pi_{hpp|i} \right) \left(\hat{\epsilon}_{lih1} / \pi_{hp|i} - \hat{\epsilon}_{lih1} / \pi_{hp|i} \right) \left(\hat{\epsilon}_{lihT} / \pi_{hp|i} - \hat{\epsilon}_{lihT} / \pi_{hp|i} \right) + \sum_{p=1}^{n_h} \hat{\Lambda}_{lih1T} / \pi_{hp|i}$$

where

$$\hat{\Lambda}_{lihp1T} = \frac{1}{2} \sum_{q=1}^{n_{hp}} \sum_{q'=1}^{n_{hp}} \left(\left(\pi_{q|ihp} \pi_{q'|ihp} - \pi_{qq'|ihp} \right) / \pi_{qq'|ihp} \right) \left(\hat{\epsilon}_{lihpq1} / \pi_{q|ihp} - \hat{\epsilon}_{lihp'q1} / \pi_{q'|ihp'} \right) \left(\hat{\epsilon}_{lihpqT} / \pi_{q|ihp} - \hat{\epsilon}_{lihp'qT} / \pi_{q'|ihp'} \right)$$

. For large samples, the standardized estimator $(\Delta \hat{Y}_{li} - \Delta Y_{li}) / \sqrt{\hat{V} \Delta \hat{Y}_{li}} \rightarrow N(0,1)$ converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for ΔY_{li} is $\Delta \hat{Y}_{li} \pm 1.96 \sqrt{\hat{V} \Delta \hat{Y}_{li}}$.

At the habitat level, the change between monitoring occasions $t=1$ and $t=T$ is estimated by, $\Delta \hat{Y}_l = \hat{Y}_{lT} - \hat{Y}_{l1}$, where $\hat{Y}_{li}$ is defined in section 3.2.1.1 for $t=1, T$. Its sampling variance is estimated by $\hat{V} \Delta \hat{Y}_l = \hat{V} (\hat{Y}_{lT} - \hat{Y}_{l1}) = \hat{V} \hat{Y}_{l1} + \hat{V} \hat{Y}_{lT} - 2 \hat{Cov}(\hat{Y}_{l1}, \hat{Y}_{lT})$, where $\hat{V} \hat{Y}_{li}$ is defined in section 3.2.1.1 for $t=1, T$, and

$\hat{Cov}(\hat{Y}_{l1}, \hat{Y}_{lT}) = 1/m(m-1) \sum_{i=1}^m (A \hat{\epsilon}_{li1} / A_i - \hat{\epsilon}_{li1})(A \hat{\epsilon}_{liT} / A_i - \hat{\epsilon}_{liT})$ estimates the covariance between the two occasion-specific habitat-level estimators. For large samples, the standardized estimator, $(\Delta \hat{Y}_l - \Delta Y_l) / \sqrt{\hat{V} \Delta \hat{Y}_l} \rightarrow N(0,1)$, converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for ΔY_l is $\Delta \hat{Y}_l \pm 1.96 \sqrt{\hat{V} \Delta \hat{Y}_l}$.

3.2.1.3.2 Ratios or proportions

At the habitat level, the change in a ratio or proportion is estimated as described in Section 3.1.3.2, replacing the design-based estimators in $\Delta \hat{R}_{CP} = \hat{R}_{CPT} - \hat{R}_{CP1}$ with the corresponding model-assisted estimators defined in Section 3.2.1.2. The covariance between the two occasion-specific model-assisted ratio estimators is estimated by

$$\hat{Cov}(\hat{R}_{CP1}, \hat{R}_{CPT}) = 1/\hat{Y}_{DM1} \hat{Y}_{DMT} \hat{Cov}(\hat{\epsilon}_{CP1}, \hat{\epsilon}_{CPT}), \text{ where}$$

$$\hat{Cov}(\hat{\epsilon}_{CP1}, \hat{\epsilon}_{CPT}) = 1/m(m-1) \sum_{i=1}^m (A \hat{\epsilon}_{iCP1} / A_i - \hat{\epsilon}_{iCP1})(A \hat{\epsilon}_{iCPT} / A_i - \hat{\epsilon}_{iCPT}). \text{ For large}$$

samples, the standardized estimator, $(\Delta \hat{R}_{CP} - \Delta R_{CP}) / \sqrt{\hat{V} \Delta \hat{R}_{CP}} \rightarrow N(0,1)$, converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for ΔR_{CP} is $\Delta \hat{R}_{CP} \pm 1.96 \sqrt{\hat{V} \Delta \hat{R}_{CP}}$.

3.2.2 Binomial logit models for tree status and age structure

To obtain model-assisted estimators of the total number and proportion of young trees, we use a binomial logit model. The model relates the probability that tree j in pixel p of stratum h within farm i is young, $\mu_{ihpj} = P(y_{ihpj} = 1)$, to an EO covariate vector

$\mathbf{x}_{ihpj} = \text{row}_{1 \leq l \leq 3}(x_{lihpj})$ through $\ln \mu_{ihpj} / (1 - \mu_{ihpj}) = \mathbf{x}_{ihpj} \boldsymbol{\beta}$. The covariate vector includes an intercept ($x_{1ihpj} = 1$) and the LiDAR-derived crown projection, x_{2ihpj} , and tree height, x_{3ihpj} . For each non-intercept covariate, x_{lihpj} , the corresponding coefficient β_l of the

parameter vector $\boldsymbol{\beta} = \text{col}_{1 \leq l \leq 3}(\beta_l)$ represents the change in the log-odds of a tree being young associated with a one-unit increase in that covariate, holding the other covariates constant. Consequently, $\exp(\beta_l)$, is the corresponding odds ratio $\mu_{ihpj} / (1 - \mu_{ihpj})$.

Let $\mathbf{B} = \text{col}_{1 \leq l \leq 3}(\beta_l)$ denote the finite-population maximum-likelihood coefficient vector that would be obtained from the complete set of response values y_{ihpj} and covariate vectors $\mathbf{x}_{ihpj}$ for all trees in the habitat. This finite-population quantity is unknown because the response values y_{ihpj} are observed only for trees included in the probability sample.

A design-consistent estimator $\hat{\mathbf{B}} = \text{col}_{1 \leq l \leq 3}(\hat{\beta}_l)$ of $\mathbf{B}$ is obtained from the sampled response values and their corresponding covariate vectors using the following Newton–Raphson updating expression:

$$\hat{\mathbf{B}}^{(m+1)} = \hat{\mathbf{B}}^{(m)} + \left[A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \mathbf{H}(y_{ihpj}, \boldsymbol{\beta}) / \pi_{hp|i} \right]_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}^{-1} A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \mathbf{b}(y_{ihpj}, \boldsymbol{\beta}) / \pi_{hp|i} \Big|_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}$$

, where $\mathbf{b}(y_{ihpj}, \boldsymbol{\beta}) = \text{col}_{1 \leq l \leq 3}[(y_{ihpj} - \mu_{ihpj})x_{lihpj}]$ and

$\mathbf{H}(y_{ihpj}, \boldsymbol{\beta}) = [(\mu_{ihpj} - \mu_{ihpj} \mu_{ihpj})] \otimes \mathbf{x}_{ihpj}^T \mathbf{x}_{ihpj}$. The $\hat{\mathbf{B}}$ covariance matrix is estimated by

$$\hat{\text{Var}}\hat{\mathbf{B}} = \left[A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \mathbf{H}(y_{ihpj}, \boldsymbol{\beta}) / \pi_{hp|i} \right]_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}^{-1} \quad (\text{Fuller, 2009}).$$

3.2.2.1 Totals

Given $\mathbf{x}_{ihpj} = \text{row}_{1 \leq l \leq 3}(x_{lihpj})$, the estimated probability that tree j in pixel p of stratum h within farm i is young is $\hat{\mu}_{ihpj} = \exp(\mathbf{x}_{ihpj} \hat{\mathbf{B}}) / (1 + \exp(\mathbf{x}_{ihpj} \hat{\mathbf{B}}))$. In principle, the LiDAR-derived covariate vector, $\mathbf{x}_{ihpj} = \text{row}_{1 \leq l \leq 3}(x_{lihpj})$, could be constructed for every candidate tree in the habitat, allowing the total number of young trees to be estimated using the full-population model-assisted estimator

$$\hat{Y}_{N_y} = \sum_{i=1}^M \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{j=1}^{y_{ihp}} \hat{\mu}_{ihpj} + A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} (y_{ihpj} - \hat{\mu}_{ihpj}) / \pi_{hp|i}.$$

However, computing $\mathbf{x}_{ihpj}$ for every candidate tree in the habitat may be costly and is not required under the adopted multistage sampling design. Auxiliary-data processing can instead be restricted to the farms selected at the first stage while retaining design-consistent habitat-level inference. For each selected farm, $i = 1, 2, \dots, m$, the total number of young trees is estimated using the model-assisted estimator

$$\hat{Y}_{N_y, i} = \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{j=1}^{y_{ihp}} \hat{\mu}_{ihpj} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} (y_{ihpj} - \hat{\mu}_{ihpj}) / \pi_{p|i}.$$

Under the usual regularity conditions, this farm-level model-assisted estimator is design-consistent for $Y_{N_y,i}$. Design consistency is ensured by the probability-weighted residual correction and does not require the working binomial logit model to be correctly specified, although the precision gain depends on its predictive performance (Firth & Bennett, 1998; Fuller, 2009).

A first-order Taylor linearization of the farm-level estimator about the finite-population coefficient vector gives $\hat{Y}_{N_y,i} \doteq Y_{N_y,i0} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} (y_{ihpj} - \mu_{ihpj0}) / \pi_{p|ih}$, where

$Y_{N_y,i0} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \mu_{ihpj0}$. It follows that the estimator is approximately design-unbiased for $Y_{N_y,i}$, that is, $E\hat{Y}_{N_y,i} \doteq Y_{N_y,i0} + Y_{N_y,i} - Y_{N_y,i0} = Y_{N_y,i}$. Its approximate design-based variance is estimated by $\hat{V}\hat{Y}_{N_y,i} \doteq \sum_{h=1}^H \hat{V} \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} (y_{ihpj} - \hat{\mu}_{ihpj}) / \pi_{p|ih}$, where $\hat{\epsilon}_{ihpj} = (y_{ihpj} - \hat{\mu}_{ihpj})$ and

$\hat{\epsilon}_{ihp} = \sum_{j=1}^{y_{ihp}} \hat{\epsilon}_{ihpj} / \pi_{p|ih}$. As a result,

$$\hat{V} \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \hat{\epsilon}_{ihpj} / \pi_{p|ih} = \frac{1}{2} \sum_{p=1}^{n_{ih}} \sum_{p'=1}^{n_{ih}} \left((\pi_{p|ih} \pi_{p'|ih} - \pi_{pp'|ih}) / \pi_{pp'|ih} \right) \left(\hat{\epsilon}_{ihp} / \pi_{p|ih} - \hat{\epsilon}_{ihp'} / \pi_{p'|ih} \right)^2.$$

At the habitat level, the total number of young trees is estimated by

$\hat{Y}_{N_y} = A1/m \sum_{i=1}^m \hat{Y}_{N_y,i} / A_i$. This estimator is approximately design-unbiased for Y_{N_y} . Its approximate design-based variance is estimated by

$$\hat{V}\hat{Y}_{N_y} = 1/m(m-1) \sum_{i=1}^m \left(A \hat{\epsilon}_{yi} / A_i - \hat{\epsilon}_y \right)^2 \text{ where } \hat{\epsilon}_{yi} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{y_{ihp}} \hat{\epsilon}_{ihpj} / \pi_{p|ih} \text{ with}$$

$\hat{\epsilon}_{ihpj} = (y_{ihpj} - \hat{\mu}_{ihpj})$, and $\hat{\epsilon}_y = A1/m \sum_{i=1}^m \hat{\epsilon}_{yi} / A_i$. For large samples, the standardized

estimator $(\hat{Y}_{N_y} - Y_{N_y}) / \sqrt{\hat{V}\hat{Y}_{N_y}} \rightarrow N(0,1)$ converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for Y_{N_y} is

$$\hat{Y}_{N_y} \pm 1.96 \sqrt{\hat{V}\hat{Y}_{N_y}}.$$

3.2.2.2 Ratios or proportions

At the habitat level, the proportion of young trees is estimated by $\hat{R}_{N_y} = \hat{Y}_{N_y} / \hat{Y}_N$, where $\hat{Y}_{N_y}$ is the model-assisted estimator of the total number of young trees defined in Section 3.2.2.1, and $\hat{Y}_N$ is the model-assisted estimator of the total number of actual trees in the habitat, given by $\hat{Y}_N = A1/m \sum_{i=1}^m \hat{Y}_{Ni} / A_i$. Here, $\hat{Y}_{Ni}$ denotes the estimated total number of actual trees in farm i .

For each LiDAR-detected candidate, let $\tilde{y}_{ihpj} = 1$ if the candidate is confirmed as an actual tree and $\tilde{y}_{ihpj} = 0$ otherwise. Using the same EO covariates as in the young-tree model, we specify $\ln \tilde{\mu}_{ihpj} / (1 - \tilde{\mu}_{ihpj}) = \mathbf{x}_{ihpj} \tilde{\boldsymbol{\beta}}$, where $\tilde{\mu}_{ihpj} = P(\tilde{y}_{ihpj} = 1 | \mathbf{x}_{ihpj})$. The coefficient vector $\tilde{\boldsymbol{\beta}}$ is estimated using the design-weighted Newton–Raphson procedure described above, yielding $\hat{\mu}_{ihpj} = \exp(\mathbf{x}_{ihpj} \hat{\mathbf{B}}) / (1 + \exp(\mathbf{x}_{ihpj} \hat{\mathbf{B}}))$. The farm-level total number of actual trees is then estimated by

$$\hat{Y}_{Ni} = \sum_{h=1}^H \sum_{p=1}^{N_{ih}} \sum_{j=1}^{\tilde{y}_{ihp}} \hat{\mu}_{ihpj} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{\tilde{y}_{ihp}} (\tilde{y}_{ihpj} - \hat{\mu}_{ihpj}) / \pi_{p|ih}.$$

By applying the Taylor linearization results in Section 3.2.2.1, with y_{ihpj} , μ_{ihpj} , and $\boldsymbol{\beta}$ replaced by $\tilde{y}_{ihpj}$, $\tilde{\mu}_{ihpj}$, and $\tilde{\boldsymbol{\beta}}$, respectively, $\hat{Y}_N$ is approximately design-unbiased for Y_N , and its approximate design-based variance is estimated analogously.

Using Taylor linearization, $\hat{Y}_{Ni} \doteq Y_{Ni0} + \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{\tilde{y}_{ihp}} (\tilde{y}_{ihpj} - \tilde{\mu}_{ihpj0}) / \pi_{p|ih}$ it is shown that $\hat{Y}_{Ni}$

is approximately unbiased, e.g., $E\hat{Y}_{Ni} \doteq Y_{Ni0} + Y_{Ni} - Y_{Ni0} = Y_{Ni}$. As a result, $\hat{Y}_N$ is approximately unbiased, and the estimator of its sampling variance is

$$\hat{V}\hat{Y}_N = 1/m(m-1) \sum_{i=1}^m \left(A \hat{\varepsilon}_{yi} / A_i - \hat{\varepsilon}_y \right)^2 \text{ where } \hat{\varepsilon}_{yi} = \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{\tilde{y}_{ihp}} (\tilde{y}_{ihpj} - \hat{\mu}_{ihpj}) / \pi_{p|ih} \text{ and}$$

$$\hat{\varepsilon}_y = A1/m \sum_{i=1}^m 1/A_i \sum_{h=1}^H \sum_{p=1}^{n_{ih}} \sum_{j=1}^{\tilde{y}_{ihp}} \hat{\varepsilon}_{yi}$$

Applying the ratio linearization in Section 3.2.1.2 to $\hat{R}_{N_y} = \hat{Y}_{N_y} / \hat{Y}_N$ gives

$\hat{R}_{N_y} \doteq R_{N_y} + 1/Y_N (\hat{Y}_{N_y} - R_{N_y} \hat{Y}_N)$, where $R_{N_y} = Y_{N_y} / Y_N$. Thus $\hat{R}_{N_y}$ is approximately design-unbiased, $E\hat{R}_{N_y} \doteq R_{N_y}$, and its approximate design-based variance is estimated by

$$\hat{V}\hat{R}_{N_y} = 1/\hat{Y}_N^2 \left[\hat{V}(\hat{Y}_{N_y}) - 2\hat{R}_{N_y} \hat{Cov}(\hat{Y}_{N_y}, \hat{Y}_N) + \hat{R}_{N_y}^2 \hat{V}(\hat{Y}_N) \right], \text{ where}$$

$$\hat{Cov}(\hat{Y}_{N_y}, \hat{Y}_N) = 1/m(m-1) \sum_{i=1}^m \left(A \hat{\varepsilon}_{N_y i} / A_i - \hat{\varepsilon}_{N_y} \right) \left(A \hat{\varepsilon}_{Ni} / A_i - \hat{\varepsilon}_N \right).$$

For large samples, the standardized estimator, $(\hat{R}_{N_y} - R_{N_y}) / \sqrt{\hat{V}\hat{R}_{N_y}} \rightarrow N(0,1)$, converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for R_{N_y} is $\hat{R}_{N_y} \pm 1.96 \sqrt{\hat{V}\hat{R}_{N_y}}$.

3.2.2.3 Change estimation

3.2.2.3.1 Totals

The change in the total number of young trees between monitoring occasions is estimated by $\Delta \hat{Y}_{N_y} = \hat{Y}_{N_y,T} - \hat{Y}_{N_y,1}$. Its approximate design-based variance is estimated by

$$\hat{V} \Delta \hat{Y}_{N_y} = \hat{V} \left(\hat{Y}_{N_y,T} - \hat{Y}_{N_y,1} \right) = \hat{V} \hat{Y}_{N_y,1} + \hat{V} \hat{Y}_{N_y,T} - 2 \hat{C} \text{ov} \left(\hat{Y}_{N_y,1}, \hat{Y}_{N_y,T} \right), \text{ where}$$

$$\hat{C} \text{ov} \left(\hat{Y}_{N_y,1}, \hat{Y}_{N_y,T} \right) = 1/m(m-1) \sum_{i=1}^m \left(A \hat{\varepsilon}_{N_y,i1} / A_i - \hat{\varepsilon}_{N_y,1} \right) \left(A \hat{\varepsilon}_{N_y,iT} / A_i - \hat{\varepsilon}_{N_y,T} \right) \text{ estimates the}$$

covariance between the two occasion-specific estimators. For $t = 1, T$, the estimator $\hat{Y}_{N_y,t}$, and its variance estimator $\hat{V} \hat{Y}_{N_y,t}$ are defined in section 3.2.2.1.

For large samples, the standardized estimator, $\left(\Delta \hat{Y}_{N_y} - \Delta Y_{N_y} \right) / \sqrt{\hat{V} \Delta \hat{Y}_{N_y}} \rightarrow N(0,1)$, converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for ΔY_{N_y} is $\Delta \hat{Y}_{N_y} \pm 1.96 \sqrt{\hat{V} \Delta \hat{Y}_{N_y}}$.

3.2.2.3.2 Ratios and proportions

The change in the proportion of young trees between monitoring occasion is estimated by $\Delta \hat{R}_{N_y} = \hat{R}_{N_y,T} - \hat{R}_{N_y,1}$. Its approximate design-based variance is estimated by

$$\hat{V} \Delta \hat{R}_{N_y} = \hat{V} \hat{R}_{N_y,T} + \hat{V} \hat{R}_{N_y,1} - 2 \hat{C} \text{ov} \left(\hat{R}_{N_y,1}, \hat{R}_{N_y,T} \right), \text{ where the occasion-specific proportion}$$

estimators and their variance estimators are defined in Section 3.2.2.2. The covariance between the two occasion-specific proportion estimators is estimated by

$$\hat{C} \text{ov} \left(\hat{R}_{N_y,1}, \hat{R}_{N_y,T} \right) = 1/\hat{Y}_{N1} \hat{Y}_{NT} \hat{C} \text{ov} \left(\hat{\varepsilon}_{N_y,1}, \hat{\varepsilon}_{N_y,T} \right), \text{ where}$$

$$\hat{C} \text{ov} \left(\hat{\varepsilon}_{N_y,1}, \hat{\varepsilon}_{N_y,T} \right) = 1/m(m-1) \sum_{i=1}^m \left(A \hat{\varepsilon}_{N_y,i1} / A_i - \hat{\varepsilon}_{N_y,1} \right) \left(A \hat{\varepsilon}_{N_y,iT} / A_i - \hat{\varepsilon}_{N_y,T} \right). \text{ For } t = 1, T, \text{ the}$$

expressions defining $\hat{\varepsilon}_{N_y,it}$ and $\hat{\varepsilon}_{N_y,t}$ are in section 3.2.2.2. For large samples, the standardized estimator interval converges in distribution to a standard normal random variable. Accordingly, an approximate 95% confidence interval for ΔR_{N_y} is

$$\Delta \hat{R}_{N_y} \pm 1.96 \sqrt{\hat{V} \Delta \hat{R}_{N_y}}.$$

3.2.3 Multinomial logit model for biodiversity assessment

Species richness and relative abundance within a habitat can be characterized through the distribution of individuals among species or functional groups. In this context, the multinomial logit model provides a flexible and interpretable framework for modelling species composition as a function of Earth observation (EO) auxiliary variables. This section presents the specification and design-consistent estimation of the multinomial logit model and its use within a model-assisted framework to estimate biodiversity indicators for habitat monitoring, including Shannon and Hill-number diversity indices.

Let $y_{ipjk} = 1$ if individual j in pixel p within farm i belongs to the species or functional group $k = 1, 2, \dots, K-1$, and $y_{ipjk} = 0$ otherwise. Let $\mathbf{x}_{ip} = \text{row}_{1 \leq l \leq L}(x_{lip})$ denote the vector of

Earth observation (EO) auxiliary variable values for pixel p . Conditional on $\mathbf{x}_{ip}$ and taking K as reference category, the multinomial logit model specifies the probability

$$\mu_{ipk} = P(y_{ipjk} = 1 | \mathbf{x}_{ip}) \text{ as } \mu_{ipk} = \exp(\mathbf{x}_{ip} \boldsymbol{\beta}_k) / \left(1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_{ip} \boldsymbol{\beta}_k) \right), \text{ for } k = 1, 2, \dots, K-1.$$

The effect of EO data on the odds ratio, $\mu_{ipk} / (1 - \mu_{ipk})$ are represented by the model parameters, $\boldsymbol{\beta} = \text{col}_{1 \leq k \leq K}(\boldsymbol{\beta}_k) = \text{col}_{1 \leq k \leq K} \text{col}_{1 \leq l \leq L}(\beta_{kl})$.

Let $\boldsymbol{\mu}_{ip} = \text{col}_{1 \leq k \leq K}(\mu_{ipk})$ denote the vector of model probabilities. To estimate μ_{ipk} at monitoring occasion t , it is sufficient to obtain estimates of $\boldsymbol{\beta}_k$ for $k = 1, 2, \dots, K-1$.

The probability associated with the reference category K is then given by

$$\mu_{ipK} = 1 - \sum_{k=1}^{K-1} \mu_{ipk} \text{ and is estimated by replacing } \boldsymbol{\beta}_k \text{ with their corresponding estimates for } k = 1, 2, \dots, K-1.$$

For LIS-based observations, a design-consistent estimator

$\hat{\mathbf{B}} = \text{col}_{1 \leq k \leq K-1}(\hat{\mathbf{B}}_k) = \text{col}_{1 \leq k \leq K-1} \text{col}_{1 \leq l \leq L}(\hat{\beta}_{kl})$ of $\boldsymbol{\beta}$ can be obtained iteratively using the Newton–Raphson updating expression

$$\hat{\mathbf{B}}^{(m+1)} = \hat{\mathbf{B}}^{(m)} + \left[A1/m \sum_{i=1}^m 1/A_i \sum_{p=1}^{n_i} \sum_{j=1}^{n_p} \mathbf{H}(\mathbf{y}_{ipj}, \boldsymbol{\beta}) / \pi_{j|i} \right]_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}^{-1} A1/m \sum_{i=1}^m 1/A_i \sum_{p=1}^{n_i} \sum_{j=1}^{n_p} \mathbf{b}(\mathbf{y}_{ipj}, \boldsymbol{\beta}) / \pi_{j|i} \Big|_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}$$

, where $\mathbf{b}(\mathbf{y}_{ipj}, \boldsymbol{\beta}) = \text{col}_{1 \leq k \leq K-1} \text{col}_{1 \leq l \leq L}[(y_{ipjk} - \mu_{ipk})x_{lip}]$, $\mathbf{y}_{ipj} = \text{col}_{1 \leq k \leq K}(y_{ipjk})$, and

$$\mathbf{H}(\mathbf{y}_{ipj}, \boldsymbol{\beta}) = \left[\text{col}_{1 \leq k \leq K-1} \text{row}_{1 \leq k' \leq K-1}(\delta_{kk'} \mu_{ipk} - \mu_{ipk} \mu_{ipk'}) \right] \otimes \mathbf{x}_{ip}^T \mathbf{x}_{ip}, \text{ with } \delta_{kk'} = 1 \text{ if } k = k', \delta_{kk'} = 0$$

otherwise (Fuller, 2009), and $\pi_{j|i} = 2L_i r_{ij} / A_i$. The same updating expression applies to third-stage quadrat sampling after setting, $\mathbf{x}_{ipq} = \mathbf{x}_{ip}$. In this case, the survey variable

is $y_{ipqk} = \sum_{j=1}^{N_{ipq}} y_{ipjk}$, representing, for example, the number of individuals belonging to

species k or to a functional group such as grasses, legumes, or forbs, and the corresponding inclusion probability is $\pi_{pq|i}$, rather than $\pi_{j|i}$.

3.2.3.1 Species probability estimates

The estimate species or functional-group probabilities for each pixel, $\hat{\boldsymbol{\mu}}_{ip} = \text{col}_{1 \leq k \leq K}(\hat{\mu}_{ipk})$,

that are compatible with the observed data $\mathbf{y}_n = \text{col}_{1 \leq i \leq m} \text{col}_{1 \leq p \leq n_i} \text{col}_{1 \leq j \leq n_p} \text{col}_{1 \leq k \leq K}(y_{ipjk})$ and maximize

the overall entropy $H(\boldsymbol{\mu}) = -\boldsymbol{\mu}^T \ln \boldsymbol{\mu} = -\sum_{i=1}^m \sum_{p=1}^{n_i} \sum_{k=1}^K \mu_{ipk} \ln \mu_{ipk}$ where

$$\boldsymbol{\mu} = \text{col}_{1 \leq i \leq m} \text{col}_{1 \leq p \leq n_i} \text{col}_{1 \leq k \leq K} (\hat{\mu}_{ipk}), \text{ are } \hat{\mu}_{ipk} = \exp(\mathbf{x}_{ip} \hat{\mathbf{B}}_k) / \left(1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_{ip} \hat{\mathbf{B}}_k) \right) \text{ for}$$

$k = 1, 2, \dots, K-1$, and $\hat{\mu}_{ipK} = 1 - \sum_{k=1}^{K-1} \hat{\mu}_{ipk} = 1 / \left(1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_{ip} \hat{\mathbf{B}}_k) \right)$ for K . Conditional on the EO auxiliary data, these estimated probabilities $\hat{\boldsymbol{\mu}}_{ip}$ are therefore fully determined by the model-parameter estimates $\hat{\mathbf{B}}$ obtained in the preceding step (Ambrosio et al., 2023).

The estimator $\hat{\boldsymbol{\mu}}_{ip} = \text{col}_{1 \leq k \leq K-1} (\hat{\mu}_{ipk})$ is design consistent. An estimator of its sampling covariance matrix, denoted by $\hat{Var}(\hat{\boldsymbol{\mu}}_{ip})$, is obtained by first-order Taylor linearization (Wolter, 2007) as $\hat{Var}(\hat{\boldsymbol{\mu}}_{ip}) \approx \hat{\mathbf{D}}_{ip} \hat{Var} \hat{\mathbf{B}} \hat{\mathbf{D}}_{ip}^T$, where $\hat{\mathbf{D}}_{ip} = \partial \hat{\boldsymbol{\mu}}_{ip} / \partial \hat{\mathbf{B}}$ is the $(K-1) \times L$ Jacobian matrix whose generic row is $\partial \hat{\mu}_{ipk} / \partial \mathbf{B}_k = \hat{\mu}_{ipk} (1 - \hat{\mu}_{ipk}) \mathbf{x}_{ip}$, and $\hat{Var} \hat{\mathbf{B}}$ is the estimator of the $(K-1)L \times (K-1)L$ sampling covariance matrix of $\hat{\mathbf{B}}$, given by

$$\hat{Var} \hat{\mathbf{B}} = \left[A1/m \sum_{i=1}^m 1/A_i \sum_{p=1}^{n_i} \sum_{j=1}^{n_{ip}} \mathbf{H}(\mathbf{y}_{ipj}, \boldsymbol{\beta}) / \pi_{j|i} \right]_{\boldsymbol{\beta}=\hat{\mathbf{B}}^{(m)}}^{-1}.$$

3.2.3.3 Shannon diversity indices

At the local level, represented here by the pixel, the Shannon diversity index is estimated as the Shannon entropy (Shannon, 1948a, 1948b),

$$H(\boldsymbol{\mu}_{ip}) = -\boldsymbol{\mu}_{ip}^T \ln \boldsymbol{\mu}_{ip} = -\sum_{k=1}^K \mu_{ipk} \ln \mu_{ipk}. \text{ At the habitat level, the pixel-level estimates are}$$

combined to obtain alpha, gamma, and beta diversity. Alpha diversity represents the mean local diversity within the habitat. It is estimated as the average of the pixel-level

Shannon indices across the habitat, $H_\alpha = 1/N \sum_{i=1}^M \sum_{p=1}^{N_i} H(\boldsymbol{\mu}_{ip})$, where N_i is the number of

pixels in farm i and $N = \sum_{i=1}^M N_i$ is the total number of pixels in the habitat.

Gamma diversity represents the diversity of the habitat considered as a whole and is estimated as the Shannon entropy calculated from the mean species or functional-group

$$\text{probabilities across all pixels, } H_\gamma = -\bar{\boldsymbol{\mu}}^T \ln \bar{\boldsymbol{\mu}} = -\sum_{k=1}^K \bar{\mu}_k \ln \bar{\mu}_k, \text{ where } \bar{\boldsymbol{\mu}} = 1/N \sum_{i=1}^M \sum_{p=1}^{N_i} \boldsymbol{\mu}_{ip}.$$

Beta diversity represents compositional heterogeneity among pixels and is estimated as $H_\beta = H_\gamma - H_\alpha$. It therefore measures the increase in Shannon entropy from the mean within-pixel diversity to the diversity of the habitat considered as a whole (Lande, 1996; De'ath, 2012).

EO auxiliary data are used to obtain model-assisted predictions of $\mathbf{\mu}_{ip}$ for every pixel p within each first-stage sampled farm $i = 1, 2, \dots, m$. These predictions are obtained from $\hat{\mathbf{\mu}}_{ip}$ for all pixels belonging to the m sampled farms. Habitat-level alpha, gamma, and beta diversity are then estimated by expanding the corresponding farm-level contributions using the first-stage inclusion probabilities. Because all pixels have the same area, the alpha-diversity estimator is $\hat{H}_\alpha = 1/m \sum_{i=1}^m 1/N_i \sum_{p=1}^{N_i} H(\hat{\mathbf{\mu}}_{ip})$, the gamma

diversity estimator is $\hat{H}_\gamma = -\hat{\mathbf{\mu}}^T \ln \hat{\mathbf{\mu}} = -\sum_{k=1}^K \hat{\mu}_k \ln \hat{\mu}_k$ with

$\hat{\mathbf{\mu}} = 1/m \sum_{i=1}^m 1/N_i \sum_{p=1}^{N_i} \hat{\mathbf{\mu}}_{ip} = \text{col} \left(\hat{\mu}_k = 1/m \sum_{i=1}^m 1/N_i \sum_{p=1}^{N_i} \hat{\mu}_{ipk} \right)$, and the beta diversity estimator is $\hat{H}_\beta = \hat{H}_\gamma - \hat{H}_\alpha$.

A design-based estimator of the sampling variance of the alpha-diversity estimator is

$\hat{V}\hat{H}_\alpha = 1/m(m-1) \sum_{i=1}^m \left(1/N_i \sum_{p=1}^{N_i} H(\hat{\mathbf{\mu}}_{ip}) - \hat{H}_\alpha \right)^2$. Because the gamma-diversity estimator,

$\hat{H}_\gamma = -\hat{\mathbf{\mu}}^T \ln \hat{\mathbf{\mu}} = -\sum_{k=1}^K \hat{\mu}_k \ln \hat{\mu}_k$, is a nonlinear function of the estimated mean

probabilities, its sampling variance is approximated by first-order Taylor linearization (Wolter, 2007) as $\hat{V}\hat{H}_\gamma = \hat{V} \sum_{k=1}^K \hat{\mu}_k \ln \hat{\mu}_k = \mathbf{d} \hat{V} \text{ar} \hat{\mathbf{\mu}} \mathbf{d}^T$, where $\mathbf{d}$ is the $1 \times K$ gradient vector

$\mathbf{d} = \partial \hat{H}_\gamma / \partial \hat{\mathbf{\mu}} = -\sum_{k=1}^K \partial \hat{\mu}_k \ln \hat{\mu}_k / \partial \hat{\mathbf{\mu}} = \left[(1 + \ln \hat{\mu}_1) \quad (1 + \ln \hat{\mu}_2) \quad \dots \quad (1 + \ln \hat{\mu}_K) \right]$, and $\hat{V} \text{ar} \hat{\mathbf{\mu}}$ is

the estimated $K \times K$ sampling covariance matrix of $\hat{\mathbf{\mu}}$. Its diagonal elements are

$\hat{V} \hat{\mu}_k = 1/m(m-1) \sum_{i=1}^m \left(1/N_i \sum_{p=1}^{N_i} \hat{\mu}_{ipk} - \hat{\mu}_k \right)^2$ for $k = 1, 2, \dots, K$, and its off-diagonal

elements are $\hat{Cov}(\hat{\mu}_k, \hat{\mu}_{k'}) = 1/m(m-1) \sum_{i=1}^m \left(1/N_i \sum_{p=1}^{N_i} \hat{\mu}_{ipk} - \hat{\mu}_k \right) \left(1/N_i \sum_{p=1}^{N_i} \hat{\mu}_{ipk'} - \hat{\mu}_{k'} \right)$ for

$k' \neq k$.

Change estimation

The change in the alpha-diversity index between monitoring occasions is estimated as $\Delta \hat{H}_\alpha = \hat{H}_{\alpha T} - \hat{H}_{\alpha 1}$. A design-based estimator of its sampling variance is

$\hat{V} \Delta \hat{H}_\alpha = \hat{V} \hat{H}_{\alpha 1} + \hat{V} \hat{H}_{\alpha T} - 2 \hat{Cov}(\hat{H}_{\alpha 1}, \hat{H}_{\alpha T})$, where

$\hat{Cov}(\hat{H}_{\alpha 1}, \hat{H}_{\alpha T}) = 1/m(m-1) \sum_{i=1}^m \left(1/N_i \sum_{p=1}^{N_i} H(\hat{\mathbf{\mu}}_{ip1}) - \hat{H}_{\alpha 1} \right) \left(1/N_i \sum_{p=1}^{N_i} H(\hat{\mathbf{\mu}}_{ipT}) - \hat{H}_{\alpha T} \right)$. For

LIS-based estimation of habitat-level alpha diversity, the covariance term includes only the first-stage covariance induced by the repeated sample of farms. The second-stage covariance is zero because the occasion-specific sets of transects are selected independently.

3.2.3.4 Hill-number diversity indices

To place Shannon entropy within a broader diversity framework, diversity can also be expressed using Hill numbers (Hill, 1973). Hill numbers constitute a continuous family of diversity measures in which the order of diversity determines the relative weight assigned to rare and abundant species or functional groups. Let $q \geq 0$ denote the order of diversity. For $q \neq 1$ the local Hill-number estimator, represented here at the pixel

level, is ${}^q\hat{D}_{ip} = \left(\sum_{k=1}^K \hat{\mu}_{ipk}^q \right)^{1/(1-q)}$. For $q = 0$, ${}^0\hat{D}_{ip} = K$ corresponds to species richness. The

case $q = 1$ is defined by continuity as ${}^1\hat{D}_{ip} = \exp \left(- \sum_{k=1}^K \hat{\mu}_{ipk} \ln \hat{\mu}_{ipk} \right)$ and represents the effective number of equally abundant species or functional groups corresponding to the

estimated Shannon entropy. For $q = 2$, ${}^2\hat{D}_{ip} = \left(\sum_{k=1}^K \hat{\mu}_{ipk}^2 \right)^{-1}$ corresponds to the inverse

Simpson index and represents the effective number of dominant species or functional groups.

At the habitat level, Hill-number diversity is estimated under the probability-sampling design by expanding the corresponding farm-level contributions using the first-stage inclusion probabilities. Specifically, for $q = 0, 1, 2$ the habitat-level Hill-number

estimator is ${}^q\hat{D} = 1/m \sum_{i=1}^m 1/N_i \sum_{p=1}^{N_i} {}^q\hat{D}_{ip}$.

A design-based estimator of its sampling variance is

$$\hat{V}({}^q\hat{D}) = 1/m(m-1) \sum_{i=1}^m \left(1/N_i \sum_{p=1}^{N_i} {}^q\hat{D}_{ip} - {}^q\hat{D} \right)^2.$$

References

- Ambrosio, L., Iglesias, L., Marín, C., & Deffense, N. (2023). Integration of remote sensing data into national statistical office sampling designs for agriculture. *Statistical Journal of the IAOS*, 39(2), 473–489. <https://doi.org/10.3233/SJI-220116>
- De'ath, G. (2012). The multinomial diversity model: Linking Shannon diversity to multiple predictors. *Ecology*, 93(10), 2286–2296. <https://doi.org/10.1890/11-2155.1>
- de Vries, P. G. (1986). *Sampling theory for forest inventory: A teach-yourself course*. Springer. <https://doi.org/10.1007/978-3-642-71581-5>
- Firth, D., & Bennett, K. E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 3–21. <https://doi.org/10.1111/1467-9868.00105>
- Fuller, W. A. (2009). *Sampling statistics*. Wiley. <https://doi.org/10.1002/9780470523551>
- Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, 54(2), 427–432. <https://doi.org/10.2307/1934352>
- Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos*, 76(1), 5–13. <https://doi.org/10.2307/3545743>
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer. <https://link.springer.com/book/9780387406206>
- Shannon, C. E. (1948a). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shannon, C. E. (1948b). A mathematical theory of communication. *Bell System Technical Journal*, 27(4), 623–656. <https://doi.org/10.1002/j.1538-7305.1948.tb00917.x>
- Thompson, S. K. (2012). *Sampling* (3rd ed.). Wiley. <https://doi.org/10.1002/9781118162934>
- Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-35099-8>