Combining Environmental Area Frame Surveys of a Finite Population

Wilmer Prentius, Xin Zhao, Anton Grafström

Wilmer Prentius

Dept. of Forest Resource Management, Swedish University of Agricultural Sciences

2019–06–19

Outline

Introduction

Goal: Produce efficient estimates for environmental surveys

Introduction

Goal: Produce efficient estimates for environmental surveys

- Motivation:
  (1) Using data from several surveys (with different goals)
  (2) Using data from a national survey together with complementary domain surveys (or vice verse)

Introduction

Goal: Produce efficient estimates for environmental surveys

- Motivation:
  (1) Using data from several surveys (with different goals)
  (2) Using data from a national survey together with complementary domain surveys (or vice verse)
- Problems:
  (1) Unknown population $\Rightarrow$ usage of area frames
  (2) Skewed data distributions (on the area frame)

U = Population
S_i = Number of inclusions of object $i \in U$ in the sample
y_i = Variable of interest measured on object i
$\pi_i = \Pr(S_i > 0)$ = inclusion probability of object i
$E_i = \mathrm{E}(S_i)$ = expected number of inclusions of object i

Unbiased if $\forall i \in U, \pi_i > 0 \ (\Rightarrow E_i > 0)$

Single-count estimator

$$Y_{SC} = \sum_{i \in U} \frac{y_i}{\pi_i} I_{S_i > 0}$$

Multiple-count estimator

$$Y_{MC} = \sum_{i \in U} \frac{y_i}{E_i} S_i$$

- Most (environmental) studies have no well defined list frames (farms, trees)

Introduction

The continuous population

- Most (environmental) studies have no well defined list frames (farms, trees)
- We have maps (i.e. a continuous population)

Introduction
The continuous population

- Most (environmental) studies have no well defined list frames (farms, trees)
- We have maps (i.e. a continuous population)
- The continuous approach requires smoothing in some way

- Most (environmental) studies have no well defined list frames (farms, trees)
- We have maps (i.e. a continuous population)
- The continuous approach requires smoothing in some way
- Calculate sample properties after the fact

Sample properties for objects

An object $i \in U$ has an inclusion zone $A_i^{(k)}$ associated with sample point $\mathbb{X}^{(k)}$. The sample point has density function $f^{(k)}$
A design P is a set of sample points.

$$S_i^{(k)} := I_{\mathbb{X}^{(k)} \in A_i^{(k)}}, \qquad S_i^{(P)} := \sum_{k \in P} I_{\mathbb{X}^{(k)} \in A_i^{(k)}},$$

$$\pi_i^{(\cdot)} := \Pr\left(S_i^{(\cdot)} > 0\right), \qquad \pi_i^{(k)} = \int_{A_i^{(k)}} f(\mathbf{x}) d\mathbf{x}, \qquad \pi_i^{(P)} = 1 - \prod_{k \in P} \left(1 - \pi_i^{(k)}\right),$$

$$E_i^{(\cdot)} := \mathrm{E}\left(S_i^{(\cdot)}\right), \qquad E_i^{(k)} = \pi_i^{(k)}, \qquad E_i^{(P)} = \sum_{k \in P} E_i^{(k)}.$$

Two totals $Y_*^{(P_1)}, Y_*^{(P_2)}, * \in \{MC, SC\}$, $\mathcal{D} = \{P_1, P_2\}$

$$Y_{L*}^{(\mathcal{D})} = \hat{\alpha} Y_*^{(P_1)} + (1 - \hat{\alpha}) Y_*^{(P_2)}, \qquad \hat{\alpha} = \frac{\hat{\mathrm{V}}\left(Y_*^{(P_2)}\right)}{\hat{\mathrm{V}}\left(Y_*^{(P_1)}\right) + \hat{\mathrm{V}}\left(Y_*^{(P_2)}\right)}$$

- When Y_* has a skewed distribution, Y_* and $\hat{\mathrm{V}}(Y_*)$ will be correlated, and the linear combination will be biased.

Two totals $Y_*^{(P_1)}, Y_*^{(P_2)}, * \in \{MC, SC\}$, $\mathcal{D} = \{P_1, P_2\}$

$$Y_{L*}^{(\mathcal{D})} = \hat{\alpha} Y_*^{(P_1)} + (1 - \hat{\alpha}) Y_*^{(P_2)}, \qquad \hat{\alpha} = \frac{\hat{\mathrm{V}}\left(Y_*^{(P_2)}\right)}{\hat{\mathrm{V}}\left(Y_*^{(P_1)}\right) + \hat{\mathrm{V}}\left(Y_*^{(P_2)}\right)}$$

- When Y_* has a skewed distribution, Y_* and $\hat{\mathrm{V}}(Y_*)$ will be correlated, and the linear combination will be biased.
- In environmental surveys high occurrence of skewedly distributed Y_*'s.

Outline

Solutions

Use the additional information to estimate variance

$$\hat{\mathrm{V}}\left(Y_{SC}^{(P_1)}\right) = \sum_{i\in U}\sum_{j\in U}\frac{y_i}{\pi_i^{(P_1)}}\frac{y_i}{\pi_j^{(P_1)}}\left(\pi_{ij}^{(P_1)} - \pi_i^{(P_1)}\pi_j^{(P_1)}\right)\frac{I_{S_i^{(P_1)}>0}I_{S_j^{(P_1)}>0}}{\pi_{ij}^{(P_1)}}$$

$\pi_i^{(P_1)} > 0\,\forall i \in U, \qquad \pi_{ij}^{(P_1)} > 0\,\forall\{i,j\} \in U,$

Solutions

Use the additional information to estimate variance

$$\hat{\mathrm{V}}\left(Y_{SC}^{(P_1)}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i^{(P_1)}} \frac{y_i}{\pi_j^{(P_1)}} \left(\pi_{ij}^{(P_1)} - \pi_i^{(P_1)} \pi_j^{(P_1)}\right) \frac{I_{S_i^{(P_1)}>0} I_{S_j^{(P_1)}>0}}{\pi_{ij}^{(P_1)}}$$

Sum over all pairs of objects in either sample

$$\hat{\mathrm{V}}_{\mathrm{LP}}\left(Y_{SC}^{(P_1)}\right) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i^{(P_1)}} \frac{y_i}{\pi_j^{(P_1)}} \left(\pi_{ij}^{(P_1)} - \pi_i^{(P_1)} \pi_j^{(P_1)}\right) \frac{I_{S_i^{(\mathcal{D})}>0} I_{S_j^{(\mathcal{D})}>0}}{\pi_{ij}^{(\mathcal{D})}}$$

$$Y_{LPSC}^{(\mathcal{D})} = \hat{\alpha}_{pool} Y_{SC}^{(P_1)} + (1 - \hat{\alpha}_{pool}) Y_{SC}^{(P_2)} \qquad \hat{\alpha}_{pool} = \frac{\hat{\mathrm{V}}_{\mathrm{LP}}\left(Y_{SC}^{(P_2)}\right)}{\hat{\mathrm{V}}_{\mathrm{LP}}\left(Y_{SC}^{(P_1)}\right) + \hat{\mathrm{V}}_{\mathrm{LP}}\left(Y_{SC}^{(P_2)}\right)}$$

$$\pi_i^{(P_1)} > 0 \, \forall i \in U, \qquad \pi_{ij}^{(P_1)} > 0 \, \forall \{i,j\} \in U, \qquad \pi_{ij}^{(\mathcal{D})} > 0 \, \forall \{i,j\} \in U, \qquad \mathcal{D} = \{P_1, P_2\}$$

Solutions

Combine the samples - sample properties for the combined design $\mathcal{D}$

$$Y_{SC}^{(\mathcal{D})} = \sum_{i \in U} \frac{y_i}{\pi_i^{(\mathcal{D})}} I_{S_i^{(\mathcal{D})} > 0} \qquad Y_{MC}^{(\mathcal{D})} = \sum_{i \in U} \frac{y_i}{E_i^{(\mathcal{D})}} S_i^{(\mathcal{D})}$$

Sample properties for objects in a set of designs $\mathcal{D} = \{P_d\}_d$

$$S_i^{(k)} := I_{\mathbb{X}^{(k)} \in A_i^{(k)}}, \quad S_i^{(P)} := \sum_{k \in P} S_i^{(k)}, \quad S_i^{(\mathcal{D})} := \sum_{P_d \in \mathcal{D}} S_i^{(P_d)},$$

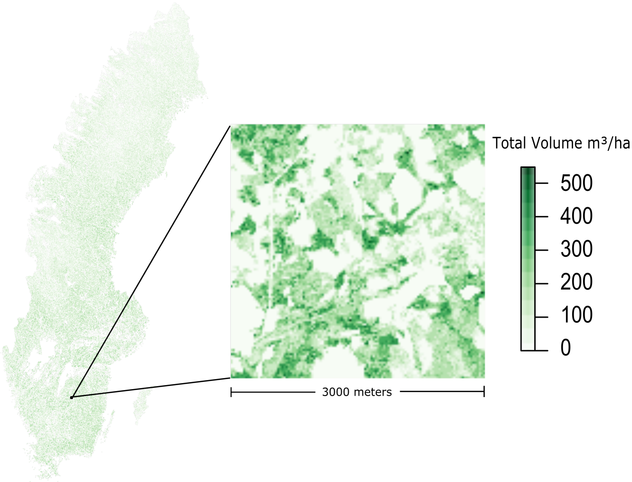
$$\pi_i^{(\cdot)} := \Pr\left(S_i^{(\cdot)} > 0\right), \quad \pi_i^{(k)} = \int_{A_i^{(k)}} f(\mathbf{x}) d\mathbf{x}, \quad \pi_i^{(P)} = 1 - \prod_{k \in P} \left(1 - \pi_i^{(k)}\right), \quad \pi_i^{(\mathcal{D})} = 1 - \prod_{P_d \in \mathcal{D}} \left(1 - \pi_i^{(P_d)}\right),$$

$$E_i^{(\cdot)} := \mathrm{E}\left(S_i^{(\cdot)}\right), \quad E_i^{(k)} = \pi_i^{(k)}, \quad E_i^{(P)} = \sum_{k \in P} E_i^{(k)}, \quad E_i^{(\mathcal{D})} = \sum_{P_d \in \mathcal{D}} E_i^{(P_d)}.$$

Outline

Simulation

Swedish Forest Map



+ individual tree data from the Swedish NFI

Simulation
Results

- Results from 10 000 simulations
- Combining two different i.i.d. designs (the worst performing case)

- Results from 10 000 simulations
- Combining two different i.i.d. designs (the worst performing case)

Empirical relative bias

- Linear combination weighted by estimated variances: -8.65 %
- Linear combination weighted by pooled estimated variances (SC): -3.73 %
- Combined sample: Unbiased

- Results from 10 000 simulations
- Combining two different i.i.d. designs (the worst performing case)

Empirical relative bias

- Linear combination weighted by estimated variances: -8.65 %
- Linear combination weighted by pooled estimated variances (SC): -3.73 %
- Combined sample: Unbiased

Reduction in MSE, compared to linear combination weighted by estimated variances:

- Linear combination weighted by pooled estimated variances (SC): 79 %
- Combined sample: 85 %

Outline

Conclusions

- Combining samples a safe bet – pooled variance estimation the efficient one
- Both will be useful for domain estimations
- Need to compute additional sample properties – easy/hard depending on setting
- Linear combinations based on estimated variances might be difficult for certain designs
- In area frame settings, sample properties depend on (accurate) positioning
- Object matching might be important

References

Grafström, A., Ekström, M., Jonsson, B.G., Esseen, P.-A., & Ståhl, G. (2019). On combining independent probability samples.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S. P., & Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics*, 28(8), e2480.

Horvitz, D., & Thompson, D. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260), 663-685.

Hansen, M. H., & Hurwitz, W. N. (1943). On the Theory of Sampling from Finite Populations. *The Annals of Mathematical Statistics*, 14(4), 333-362.

Pictures:
https://www.skogsstyrelsen.se/sjalvservice/karttjanster/skogliga-grunddata/ (Skogsstyrelsen)
https://www.slu.se/centrumbildningar-och-projekt/nils/Datainsamling/faltinventering/ (NILS)
SLU Forest Map:
https://www.slu.se/en/Collaborative-Centres-and-Projects/the-swedish-national-forest-inventory/forest-statistics/slu-forest-map/