A new variance estimator for spatially restricted sampling designs

Based on Continuous Horwitz-Thompson theorem (CHT)

Radim Adolt

Forest Management Institute Brandýs nad Labem, branch Kroměříž
National Forest Inventory Methodology and Analysis

2nd International Workshop on Forest Inventory Statistics
Kroměříž, 24th - 26th May 2016

Contents

1. Continuous Horwitz-Thompson theorem
2. Important aspects of NFI sampling
3. Derivation of the estimator and its properties
4. Comparison to an already known variance estimator

Estimation of population's total using HTC [Cordy, 1993]

Let Y be the total of a resource in geographical domain D made up by an infinite set of points x (D is a continuous population)

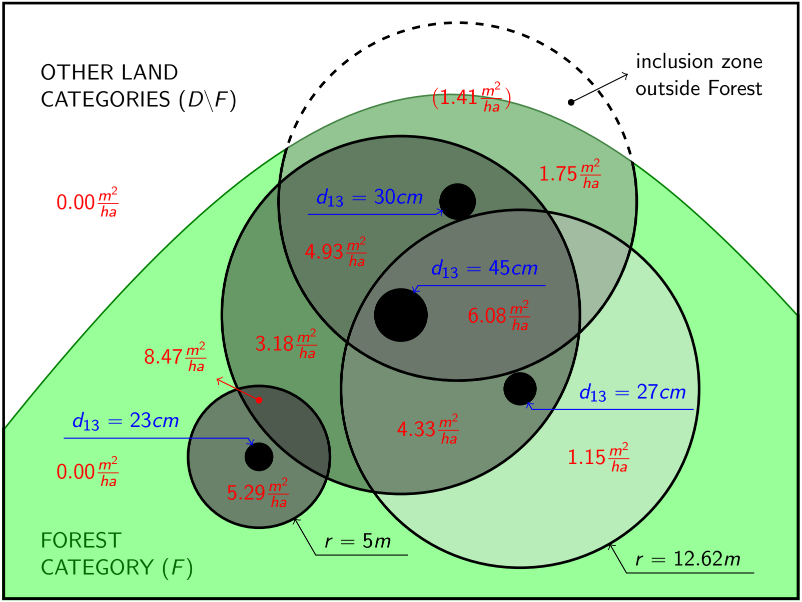
$$Y = \int_D Y(x)dx = \lambda(D)\bar{Y}. \quad (1)$$

An unbiased estimator $\hat{Y}$ **of the total** Y is given by

$$\hat{Y} = \sum_{x \in s} \frac{Y(x)}{\pi(x)}, \quad (2)$$

where the symbol $Y(x)$ stands for local density found on the point x of the sample s with fixed size n_D, $\pi(x)$ is the **inclusion density function** on the point x (local measure of the expected number of sample points per unit area).
Estimator (2) is unbiased if the function $Y(x)$ is positive or bounded in D and if $\pi(x) > 0$ $\forall x \in D$, i. e. any point $x \in D$ can be selected.

STUDY AREA
GEOGRAPHICAL DOMAIN (D)



Variance of total estimator [Cordy, 1993]

If the local density $Y(x)$ is bounded, conditions $\pi(x) > 0 \ \forall x \in \ D$ and $\int_D \frac{1}{\pi(x)} dx < \ \infty$ hold, the **variance of total estimator** $\hat{Y}$ is given:

$$\mathbb{V}_{HTC}\big(\hat{Y}\big) = \int_D \frac{Y^2(x)}{\pi(x)} dx + \\ + \int_D \int_D Y(x_i) Y(x_j) \left[\frac{\pi(x_i, x_j) - \pi(x_i)\pi(x_j)}{\pi(x_i)\pi(x_j)} \right] dx_i dx_j. \quad (3)$$

Estimation of the variance of total estimator [Cordy, 1993]

Provided condition $\pi(x_i, x_j) > 0 \; \forall x_i, x_j \in D$ is satisfied, an **unbiased estimator of variance** $\mathbb{V}(\hat{Y})$ is given by

$$\hat{\mathbb{V}}_{HTC}(\hat{Y}) = \sum_{x_i \in s} \left[\frac{Y(x_i)}{\pi(x_i)} \right]^2 + \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} Y(x_i) Y(x_j) \left[\frac{\pi(x_i, x_j) - \pi(x_i)\pi(x_j)}{\pi(x_i, x_j)\pi(x_i)\pi(x_j)} \right], \quad (4)$$

and equivalently by

$$\hat{\mathbb{V}}_{HTC}(\hat{Y}) = \sum_{x_i \in s} \left[\frac{Y(x_i)}{\pi(x_i)} \right]^2 + \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} \frac{Y(x_i) Y(x_j)}{\pi(x_i)\pi(x_j)} - \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} \frac{Y(x_i) Y(x_j)}{\pi(x_i, x_j)}. \quad (5)$$

Total and variance estimators for URS

In case of **Uniform Random Sampling (URS)** with a **fixed number of n_D sample points selected in geographical domain** D the general HTC estimators (2) and (4), (5) take the form of (6), (7), (8):

$$\hat{Y} = \sum_{x \in \mathbf{s}} \frac{Y(x)}{\pi(x)} = \lambda(D) \frac{\sum_{x \in \mathbf{s}} Y(x)}{n_D} = \lambda(D) \hat{\bar{Y}}, \tag{6}$$

$$\hat{\mathbb{V}}_{URS}(\hat{Y}) = \frac{\lambda^2(D)}{n_D(n_D - 1)} \sum_{x \in \mathbf{s}} \left[Y(x) - \hat{\bar{Y}} \right]^2, \tag{7}$$

$$\hat{\mathbb{V}}_{URS}(\hat{Y}) = \frac{\lambda^2(D)}{n_D - 1} \left[\widehat{\overline{Y^2}} - \hat{\bar{Y}}^2 \right]. \tag{8}$$

Problems with NFI variance estimation I

NFI sampling design properties influencing variance estimation

- most NFIs use a variant of **systematic** or **spatially stratified design**
- **condition** $\pi(x_i, x_j) > 0 \ \forall x_i, x_j \in D$ **doesn't hold** for systematic sampling and for stratified sampling with fixed origin

There are several workarounds how to approximate and estimate the design-based variance of systematic samples see Wolter M. [1985], Cordy and Thompson [1995], Heikkinen [2006, section 10 starting on page 155] and Cooper [2006] for instance.

Problems with NFI variance estimation II

Variance estimators based on URS approximation

- for designs for which $\pi(x_i, x_j) > 0 \ \forall x_i, x_j \in D$ doesn't hold, **an approximation of** $\pi(x_i, x_j) = n_D(n_D - 1)/\lambda^2(D)$ **corresponding to URS with fixed sample size in** D is used
- **variance estimators are approximated** by (7), (8)
- under the condition of **positive spatial correlation** and sufficient sample size **these approximations of variance estimator are mostly but not necessarily**[a] **conservative** [Heikkinen, 2006, section 10 starting on page 155] and [Mandallaz, 2007, section 4.1 starting on page 53]

---

[a] Sampling grid resolution might coincide with periodicity of the local density function.

Conservativeness of the variance estimator

Definition

$$\mathbb{E}\Big[\hat{\mathbb{V}}(\hat{\theta})\Big] > \mathbb{V}\big(\hat{Y}\big) \tag{9}$$

Expected value of **a conservative variance estimator** $\hat{\mathbb{V}}\big(\hat{\theta}\big)$ is higher than the true variance.

Non-constant sample size in D I

Reasons for and implications of non-constant sample size

- most NFIs use square, less frequently rectangular, triangular or hexagonal sampling grids with random origin
- sampling grid divide the **the support**[a] $\mathcal{S} \supseteq D$ into congruent, non-overlapping cells of equal size and shape - **inventory blocks**
- D **can't be tessellated by a whole number of inventory blocks**
- because of the random origin of the grid and for some designs also due to random location of sample points in inventory blocks, **the number of sample points in** D **varies from sample to sample**
- **HTC formulas in the form given at the begining of presentation do not hold for random sample size designs!**

[a] An area used for the implementation of sampling algorithm - typically a rectangle covering D.

Non-constant sample size in D II

Reaching unbiasedness of estimators under non-constant sample size

One option is **to implement the estimation on $\mathcal{S}$ instead of D** level[a] and to redefine local density to zero in areas outside D. This technique **works well for designs not using the URS approximation for variance estimation**, otherwise it further increases the conservativeness of variance estimator.

Cordy [1993] describes two solutions:

1. replace inclusion density and pairwise inclusion density functions on inventory points by their expected values taken over all possible sample sizes, estimators and their variances are then **unbiased unconditionally on the realized sample size in D, total estimators are additive**
2. express the inclusion density and pairwise inclusion density functions with respect to realized sample size in D, estimators are **unbiased conditionally on the realized sample size in D, total estimators are no longer additive**

---

[a] Specific topology for $\mathcal{S}$ [Stevens, 1997, section 3.1 starting on page 172], [Mandallaz, 2007, section 5.6 starting on page 92] the number of sample points in $\mathcal{S}$ is kept constant.

Geographical additivity of total estimators

Geographical additivity of total estimator $\hat{Y}$ is defined by

$$\hat{Y}_{\bigcup_{i=1}^{k} D_i} = \sum_{i=1}^{k} \hat{Y}_{D_i}. \tag{10}$$

A geographically additive total estimator evaluated for arbitrary domain being an union $\bigcup_{i=1}^{k} D_i$ of k sub-domains $D_1, D_2 \ldots D_k$ equals to the sum of k total estimators of the same kind calculated individually for each sub-domain.

Unconditional versus conditional unbiasedness of estimators

Under **unconditional unbiasedness** the expected value of an estimator $\hat{\theta}$, **when evaluated over all possible sample sizes**, equals the true value of the parameter of given population

$$\mathbb{E}\left[\hat{\theta}\right] = \theta. \tag{11}$$

Under **conditional unbiasedness** (conditioning on sample size) the expected value of an estimator $\hat{\theta}$, **when evaluated over all samples of given size** n_D, equals the true value of the parameter of given population

$$\mathbb{E}\left[\hat{\theta} \mid n_D\right] = \theta. \tag{12}$$

Unconditionally unbiased, additive estimator of total I

Cordy's [1993] first option how to cope with random sample size - **replace inclusion density and pairwise inclusion density functions on inventory points by their expected values:**

$$\bar{\pi}(x) = \mathbb{E}\big[\pi_{n_D}(x)\big] = \sum \alpha_{n_D}\pi_n(x), \tag{13}$$

$$\bar{\pi}(x_i, x_j) = \mathbb{E}\big[\pi_{n_D}(x_i, x_j)\big] = \sum \alpha_{n_D}\pi_n(x_i, x_j). \tag{14}$$

Sums in formulas (13) and (14) run over all possible sample sizes, α_{n_D} is a probability of obtaining a sample of size n_D.

Skipping a rather trivial algebra the expected value of sampling density on location x can be intuitively expressed by

$$\bar{\pi}(x) = \lambda(c)^{-1}, \tag{15}$$

being a reciprocal value of the area of one inventory block (e.g. square).

Unconditionally unbiased, additive estimator of total II

Derivation of $\bar{\pi}(x_i, x_j)$ is based on **the assumption that for every sample of particular size n_D the pairwise density can be approximated by** $\pi_{n_D}(x_i, x_j) \approx n_D(n_D - 1)/\lambda^2(D)$ (the URS case). Then the expected value $\bar{\pi}(x_i, x_j)$ (over all possible sample sizes) can be expressed by:

$$\bar{\pi}(x_i, x_j) = \mathbb{E}\left[\frac{n_D(n_D - 1)}{\lambda^2(D)}\right] = \frac{\mathbb{E}(n_D^2) - \mathbb{E}(n_D)}{\lambda^2(D)} = \frac{\mathbb{E}(n_D^2) - \bar{n}_D}{\lambda^2(D)}, \quad (16)$$

where of $\bar{n}_D = \lambda(D)/\lambda(c)$ and $\mathbb{E}(n_D^2)$ can be taken from:

$$\mathbb{V}(n_D) = \mathbb{E}(n_D^2) - \mathbb{E}^2(n_D). \quad (17)$$

Putting pieces together we arrive at

$$\bar{\pi}(x_i, x_j) = \bar{n}_D(\bar{n}_D - 1 + \mathbb{V}(n_D)/\bar{n}_D)\lambda^{-2}(D). \quad (18)$$

Unconditionally unbiased, additive estimator of total III

Densities (15) and (18) can be plugged into (6), (4) or (5) to get following estimator of total - **additive and unbiased unconditionally on sample size**

$$\hat{Y} = \sum_{x \in s} \frac{Y(x)}{\bar{\pi}(x)} = \lambda(D) \frac{\sum_{x \in s} Y(x)}{\bar{n}_D} = \lambda(c) \sum_{x \in s} Y(x), \tag{19}$$

and its variance estimator

$$\hat{\mathbb{V}}\big(\hat{Y}\big) = \frac{\lambda^2(D)}{\bar{n}_D - 1 + \hat{\mathbb{V}}(n_D)/\bar{n}_D} \left\{ \widehat{Y^2} + \hat{\hat{Y}}^2 \Big[\hat{\mathbb{V}}(n_D)/\bar{n}_D - 1 \Big] \right\}. \tag{20}$$

In practice we need to replace $\mathbb{V}(n_D)$ by its estimate which can be obtained by simulated generation of sampling grid over a digital map of a study area (or by another approximate approach).

The estimators **use only sample points located in D which is computationally efficient**. The estimator of variance is conservative in most NFI situations - due to the URS approximation behind.

The already known unconditionally unbiased estimator I

Total estimator (unbiased unconditionally) can be expressed as

$$\hat{Y} = \hat{\lambda}(D)\hat{\hat{Y}}_D = n_D \lambda(c) \hat{\hat{Y}}, \tag{21}$$

where $\hat{\hat{Y}}$ is the estimator of mean density over D, and $\hat{\lambda}(D)$ is the estimator of area of D. Estimator (21) is a product of two independent random variables (n_D and $\hat{\hat{Y}}$) and a constant $\lambda(c)$ (cell size of the grid).

Goodman [1960] showed an exact variance estimator of a product of two uncorelated random variables, we use this formula to get following variance estimator (personal communication with Mr. Adrian Lanz)

$$\hat{\mathbb{V}}(\hat{Y}) = \lambda^2(c)\left[n_D \hat{\sigma}_D^2 + \hat{\hat{Y}}^2 \hat{\mathbb{V}}(n_D) - \frac{\hat{\mathbb{V}}(n_D)}{n_D}\hat{\sigma}_D^2\right], \tag{22}$$

where

$$\hat{\sigma}_D^2 = \frac{\sum_{x \in s}\left[Y(x) - \hat{\hat{Y}}\right]^2}{n_D - 1}, \tag{23}$$

is an estimator of variance of local density $Y(x)$ in D.

The already known unconditionally unbiased estimator II

Estimator (22) can be expressed as

$$\hat{\mathbb{V}}\big(\hat{Y}\big) = \lambda^2(c)\left[n_D^2\frac{\widehat{\overline{Y^2}} - \hat{\bar{Y}}^2}{n_D - 1} + \hat{\bar{Y}}^2\hat{\mathbb{V}}\big(n_D\big) - \hat{\mathbb{V}}\big(n_D\big)\frac{\widehat{\overline{Y^2}} - \hat{\bar{Y}}^2}{n_D - 1}\right], \qquad (24)$$

and further rewritten as

$$\hat{\mathbb{V}}\big(\hat{Y}\big) = \frac{\lambda^2(D)}{n_D - 1}\left\{\widehat{\overline{Y^2}}\Big[1 - \hat{\mathbb{V}}(n_D)/n_D^2\Big] + \hat{\bar{Y}}^2\Big[\hat{\mathbb{V}}(n_D)/n_D - 1\Big]\right\}. \qquad (25)$$

Now let's compare (25) to (26) derived from HTC:

$$\hat{\mathbb{V}}\big(\hat{Y}\big) = \frac{\lambda^2(D)}{\bar{n}_D - 1 + \hat{\mathbb{V}}(n_D)/\bar{n}_D}\left\{\widehat{\overline{Y^2}} + \hat{\bar{Y}}^2\Big[\hat{\mathbb{V}}(n_D)/\bar{n}_D - 1\Big]\right\}. \qquad (26)$$

Summary

- NFI estimation can be based upon HTC by Cordy [1993]
- additivity of totals can be addressed by unconditionally unbiased estimator
- for a spatially restricted designs two URS-based (conservative) variance estimators are available
- the new one was derived from HTC, the former is based on standard theory of random sampling
- in theory these estimators are almost identical which was confirmed by simulations using artificial population

References I

C. Cooper. Sampling and variance estimation on continuous domains. *Environmetrics*, 17:539–553, 2006.

B. Cordy, C. and M. Thompson, C. An application of the deterministic variogram to design-based variance estimation. *Mathematical Geology*, 27:173–205, 1995.

C. B. Cordy. An extension of the horwitz-thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, 18:353–362, 1993.

L. A. Goodman. On the exact variance of products. *Journal of the American Statistical Association*, 55:708–713, 1960.

J. (eds. Kangas A. & Maltamo M.) Heikkinen. *Forest Inventory, Methodology and Applications*, chapter Assessment of Uncetainty in Spatially Systematic Sampling, pages 155–176. Springer Verlag, 2006.

D. Mandallaz. *Sampling Techniques For Forest Inventories*. Chapman and Hall/CRC, 2007.

D. L. Jr. Stevens. Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*, 8:167–195, 1997.

K. Wolter M. *Introduction to Variance Estimation*. Springer, 1985.

End of presentation

Thank you for attention!