Deliverable

D2.3 Statistically-sound estimators

PathFinder Project

Version: 2.3

27 August 2025

The research leading to these results has received funding from the European Union Horizon Europe (HORIZON) Research & Innovation programme under the Grant Agreement no. 101056907

I. DOCUMENT CONTROL

Project	PathFinder (101056907)
Project Title	Towards an integrated consistent European LULUCF monitoring and policy pathway assessment framework
Date	27 August 2025
Author/s	Radim Adolt NLI, Czech Forestry Institute
Reviewer/s	Ambros Berger, Christoph Kleinn, Lutz Fehrman, Ryan Caroll, Johannes Breidenbach
Activity	WP2, Task 2.3 Statistically-sound estimators for combining field and remotely sensed data
Dissemination Level	PU
Filename	pathfinder_d23 _v23.pdf

DISSEMINATION LEVEL

PU Public, fully open access

II. DOCUMENT HISTORY

Version	Date	Author	Change
0.1	12 June 2025	Radim Adolt	Full draft for internal review
1.0	1 July 2025	Radim Adolt	Corrected full draft for internal review
1.1	24 July 2025	Radim Adolt	Internal review / suggestions incorporated
1.3	4 August 2025	Radim Adolt	Abbreviations and references complemented
1.4	18 August 2025	Radim Adolt	Annex C updated, section 3.4 WIP
2.0	24 August 2025	Radim Adolt	A, C, 3.2 and 3.4 updated
2.1	25 August 2025	Radim Adolt	Title page updated – CFI changed for NLI
2.2	26 August 2025	Radim Adolt	List of reviewers added, Table 3 extended
2.3	27 August 2025	Radim Adolt	Table 3 and Annex B reduced – final version

Mé ženě Božence a mým dětem Julince a Honzíčkovi,
s láskou a nadějí v srdci.

Contents

List of Abbreviations

AGB Above Ground Biomass

FMI Forest Management Inventory

HTC Horvitz-Thompson Theorem for Continuous populations [Cordy, 1993]

NFI National Forest Inventory

NUTS Nomenclature of Territorial Units for Statistics

ICP Forests The International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests

INSPIRE Infrastructure for Spatial Information in the European Community

nFIESTA new Forest Inventory and Analysis (https://gitlab.com/groups/nfiesta/-/wikis/home)

1p single-phase estimator

1pm single-phase regression estimator with exhaustive auxiliaries (a map)

2p two-phase regression estimator with auxiliaries in the first phase

2pm two-phase regression estimator with 1st-phase and exhaustive auxiliaries (a map)

1 Introduction

This report is the deliverable D2.3 of the EU Horizon PathFinder project. It is an outcome of task *'T2.3: Statistically sound estimators for combining field and remotely sensed data'*. The main subject of the report is the estimation of totals, ratios, and differences of totals or ratios, in settings that are relevant for forest and environmental monitoring. The emphasis is on the use of auxiliary data together with field data (ground truth) to increase the precision of the estimators.

In forest inventories, the auxiliary information is usually sourced from remote sensing products (digital maps). Large sets of sample plots with attributes obtained by remote sensing or from a past inventory represent another form of auxiliaries. The auxiliaries are typically cheaper and often updated with a higher frequency than the field data. However, they might be biased. The proposed methods convey the best of the two data sources in the resulting estimates, the unbiasedness of the field data, and the large amounts of information contained in the auxiliaries leading to precision improvement.

The design-based inferential framework is followed, meaning that the surveyed population is considered fixed and the only source of variation of estimates is the sample. If the sampling is repeated, different population units are likely to be selected, and the estimates change. The relationship between the field data and auxiliaries is captured by linear regression models that do not need to be correct for the estimators to remain unbiased and become more precise. This is the principle of model-assisted estimation that belongs to design-based inference [Särndal *et al.*, 2003].

The best possible total estimators for different target variables, e.g., for Above Ground Biomass (AGB) and for forest area, are often not of the same type. Instead, they depend on the availability of a particular type of auxiliaries (large first-phase samples, maps, or both). Many parameters that need to be estimated are actually defined as ratios of two totals such as AGB (the numerator) per hectare of forest land (the denominator). Changes in totals, for example, the change in AGB, or changes in ratios, e. g., the change in the AGB per hectare of forest land, are of particular importance in environmental monitoring programmes because they show potential trends. As the availability of auxiliary data changes over time, these differences are inevitably defined as mixtures of various estimator types that are optimal for particular periods only.

The methods of Adolt *et al.* [2018] and those presented in this report address the estimation of all possible single-phase (with or without exhaustive auxiliaries) and two-phase regression totals, their ratios, their differences and differences of their ratios.

The proposed estimators are generic in the sense that they are constructed to (i) work with the infinite population approach to forest inventory [Mandallaz, 1991], (ii) be applicable to designs beyond uniform random sampling, and (iii) in the case of totals and their differences, preserve geographical additivity. The first two conditions have been addressed by the Horvitz-Thompson theorem for continuous populations formulated by Cordy [1993].

Under geographical additivity, the total estimate for a region $D_+ = \bigcup_{i=1} D_i$ matches the sum of estimates of its all non-overlapping sub-regions D_i. A linear model is fitted in D_+ and then applied to the estimation cells D_i. To preserve additivity if the sample plots are organized in clusters, the cluster-level values of attributes are calculated by the weight-share method [Bouriaud *et al.*, 2024].

Last but not least, the estimators of ratios, the estimators of difference of totals, and the estimators of difference of ratios of totals are formulated in two variants depending

on whether all involved total estimators use the same sample or not. Different samples (either totally distinct or overlapping) are often the case if we evaluate ratios or differences between two periods, in which different sets of (temporary) plots or moving-window estimators of different lengths may be used.

Note that there is no point of recommending one 'best' estimator from the set proposed here. All estimators are design-based and, therefore unbiased (at least approximately), and their variance (precision) depends on whether they use any auxiliaries or not and, if so, how well these auxiliaries correlate with values of the particular variable observed in the field. So, the choice of estimator is primarily driven by its nature (total, ratio of totals, difference of totals, difference of ratios of totals) and availability of auxiliaries. Sometimes, estimators using one type of auxiliary (either first-phase auxiliaries or a map) can be more precise than an estimator integrating both types [Estevao & Särndal, 2004]. This situation can be detected by calculating both estimators and comparing their estimated variances.

Chapter 2 elaborates new two-phase regression estimators of total t_y in a geographical region of estimation D (the estimation cell).

$$t_y = \int_D y(x)dx \tag{1}$$

The first estimator uses auxiliary data available in the first phase of sampling (large set of plots, of which the field data is a subsample). The second estimator adds exhaustive auxiliaries, that is, maps from which vectors of auxiliaries for the first- and second-phase sample points are extracted, and 'error-free' totals calculated for the desired regions of estimation. If the auxiliaries in the first phase correspond or include field data measured in the past, the correlations with the current field data (second-phase sample) can be reasonably high, leading to substantial precision gains. These techniques were proposed by Massey & Lanz [2014]. However, the authors only assumed uniform random sampling of the geographical area, and their local density definition did not preserve the additivity of the totals.

Chapter 3 presents new estimators for target parameters defined as ratios of two variables. The various types of two-phase estimators in the numerator and denominator generate a series of 19 estimators that include but go far beyond those proposed by Adolt *et al.* [2018]. These estimators are needed to be able to present the best possible total estimates for the numerator, for the denominator, and an estimate of the ratio that exactly corresponds to their division. This internal consistency is important from the interpretation and end-user perspectives.

New difference estimators are formulated in the chapters 4 (differences of totals, a series of 11 estimators) and 5 (differences of ratios, a series of 261 estimators). These estimators seamlessly fill the space between direct and indirect estimators of change [McRoberts *et al.*, 2015], depending on the proportion of common panels ranging from 0% (indirect change estimators) to 100% (direct change estimators).

All estimators included in the chapters 2, 3, 4, and 5 are proposed in view of their implementation in nFIESTA, that is, in the estimation component of the PathFinder platform described in an earlier report of Miettinen *et al.* [2024].

In the last chapter 6 the implementation of selected estimators is summarised.

The addendum contains three side topics that are practically relevant but did not fit the overall structure of the report.

(i) Merging of panels (spatially representative samples) from different time periods is described in Section 6.4. Sometimes, panels from the same period are combined. For

example, temporary and permanent sample plots with a different design can be used together for estimation within an NFI (National Forest Inventory). Another example might be combination of panels from different sources such as the NFI, ICP Forests (The International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests) or FMI (Forest Management Inventory) data. In all these examples, the method of Section 6.4 is not optimal because estimates based on individual panels are likely to have different design-based variances. Therefore, Addendum A.1 proposes a method for the calculation of estimates based on various panels coming from the same period but with a different sampling design (selection of sample points) or with a different definition of attribute density (plot or cluster configuration).

(ii) If maps are produced using field sample plots as training datasets, then there is the question of whether these maps can serve as auxiliary data in model-assisted estimation using the same sample plots. In this context, the estimators use the so-called endogenous model. The approach is justified by the findings of relevant studies in Addendum A.2.

(iii) Maps produced by complex procedures, such as machine learning algorithms, may overfit, i.e., be closer to reality at the positions near training data compared to the remaining parts of the map. If such maps are used for model-assisted estimation with an endogenous model, the variance of the estimator will be underestimated. In Addendum A.3 a method is proposed to circumvent the variance underestimation.

In Annexes A and B the variances end their estimators are derived for single- and two-phase totals, and for any combination of single- and two-phase total used in a ratio. Annex C presents derivations of covariances and their estimators between any combination of single- and two-phase totals.

2 Generic two-phase regression estimators of the total

2.1 Two-phase regression estimator of the total with the first-phase auxiliaries

This type of estimator combines auxiliary variables that are available in a larger sample of plots selected in the first phase (s_1) with a relatively smaller sample of field data collected for a subset of the large sample, on the so-called second-phase plots ($s2$). If there is a substantial correlation between the auxiliaries and field data, the precision of the estimator is improved compared to using only field data.

The estimator of the total of variable y in the geographical region D (estimation cell) is defined:

$$\hat{t}_{y,2p} = \mathbf{I_D}\mathbf{\Pi_D}\mathbf{Y'_D} + \mathbf{\Delta_{z^{(1)}}}'\tilde{\mathbf{G}}_{\beta_+^{(1)}}\mathbf{\Pi}_+\mathbf{Y}'_+ \tag{2}$$

where the first term corresponds to the single-phase estimator of the total of target variable y in D, and the second term corresponds to a correction based on the difference between the vectors of auxiliary totals estimated from the first-phase (large) and from the second-phase sample ($\mathbf{\Delta_{z^{(1)}}}$) and the estimated coefficients of a linear model ($\tilde{\mathbf{G}}_{\beta_+^{(1)}}\mathbf{\Pi}_+\mathbf{Y}'_+$). The meaning of individual symbols is the following:

- $\mathbf{I_D}$ ... $1 \times n_2$ vector of indicator variables $I_D(x)$ taking a value of 1, if any plot of a cluster (or the single plot) is found within the estimation cell D and 0 otherwise, n_2 corresponds to the overall number of 2nd-phase plots,
- $\mathbf{\Pi_D}$ is the $n_2 \times n_2$ matrix with diagonal elements corresponding to inverse sampling densities $\pi^*(x)^{-1} = \pi_1(x)^{-1}\pi_{2|1}(x)^{-1}$ where $\pi_1(x)$ is the inclusion density of the first–phase [Cordy, 1993], and $\pi_{2|1}(x)$ is the sampling probability of the second phase given the sample of the first phase. The inclusion density $\pi^*(x)$ is an analogy to the selection probability π^* introduced by Särndal *et al.* [2003, p. 347] in the context of two-phase sampling.
- $\mathbf{Y_D}$ $1 \times n_2$ vector of densities at cluster level $y^{(D)}(x) = \frac{\sum_{k=1}^{m} I_{kD} y_k}{m}$, m corresponding to the nominal number of plots within a cluster and y_k corresponding to local densities at the plot level, I_{kD} is an indicator variable taking value of 1 if the centre of the plot k is located inside D,
- $\mathbf{\Delta_{z^{(1)}}} = \hat{\mathbf{t}}_{\mathbf{s_1},\mathbf{z_D^{(1)}}} - \hat{\mathbf{t}}_{\mathbf{s_2},\mathbf{z_D^{(1)}}} = \frac{\sum_{x\in_1} \mathbf{Z_D^{(1)}}(x)}{\pi_1(x)} - \frac{\sum x\in_2 \mathbf{Z_D^{(1)}}(x)}{\pi^*(x)}$ is the $p_1 \times 1$ vector of differences between the single-phase estimates of the p_1 number of auxiliaries using the first and the second-phase samples s_1 and s_2,
- $\mathbf{Z_D^{(1)}}(x) = \frac{\sum_{k=1}^{m} I_{kD}\mathbf{Z_k^{(1)}}}{m}$ where $\mathbf{Z}_k^{(1)}$ is the auxiliary vector at plot k (available at 1st-phase plots, but not everywhere in D_+), I_{kD} takes value of 1 if and only if the centre of the plot k belongs to D,

$$\tilde{\mathbf{G}}_{\beta_+^{(1)}} = \left[\mathbf{Z_+^{(1)}}\mathbf{\Sigma\Pi}_+\mathbf{Z_+^{(1)}}'\right]^{-1}\mathbf{Z_+^{(1)}}\mathbf{\Sigma}_+ \tag{3}$$

- $\mathbf{Z_+^{(1)}}$ is the $p_1 \times n_{2,+}$ matrix of (non-exhaustive) auxiliary densities available only for the 1st–phase plots, the columns of $\mathbf{Z_+^{(1)}}$ correspond to $p_1 \times 1$ auxiliary vector $\mathbf{Z_+^{(1)}}(x)$ at the sample point x, $n_{2,+}$ is the number of sample points in the parameterisation region D_+.

- $\mathbf{Z}_{+}^{(\mathbf{1})}(x) = \frac{\sum_{k=1}^{m} I_{kD_+} \mathbf{Z}_{\mathbf{k}}^{(\mathbf{1})}}{m}$ with $\mathbf{Z}_{\mathbf{k}}^{(\mathbf{1})}$ is the auxiliary vector for plot k, I_{kD_+} takes value 1 if and only if the centre of the plot k belongs to the parameterisation region D_+,
- $\mathbf{\Sigma}_+$ is an $n_{2,+} \times n_{2,+}$ matrix with diagonal elements corresponding to inverse values of the anticipated variances $\sigma^2(x)$ used for weighted least squares. A specific setting of the matrix is proposed for designs using clusters; see Adolt *et al.* [2018].
- $\mathbf{\Pi}_+$ is $n_{2,+} \times n_{2,+}$ matrix with diagonal elements corresponding to inverse sampling densities $\pi^*(x)^{-1}$
- $\mathbf{Y}_+$ ... $1 \times n_2$ vector of cluster-level densities $y^{(D+)}(x) = \frac{\sum_{k=1}^{m} I_{kD_+} y_k}{m}$ with y_k corresponding to local densities at plot level.

The estimator of variance of $\hat{t}_{y,2p}$ is given by

$$\begin{aligned}
\hat{\mathbb{V}}\left(\hat{t}_{y,2p}\right) = & \sum_{j=1}^{J} \sum_{x \in s_{2,j}} \frac{y^{(D)}(x)^2}{\pi^*(x)\pi_1(x)} + \\
& + \sum_{j=1}^{J} \sum_{x \in s_{2,j}} \sum_{\substack{x' \in s_{2,j} \\ x' \neq x'}} y^{(D)}(x) y^{(D)}(x') \frac{\pi_1(x,x') - \pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} + \\
& + \sum_{j=1}^{J} \sum_{x \in s_{2,j}} \left[\frac{\rho(x)}{\pi^*(x)}\right]^2 \left[1 - \pi_{2|1}(x)\right] + \\
& + \sum_{j=1}^{J} \sum_{x \in s_{2,j}} \sum_{\substack{x' \in s_{2,j} \\ x' \neq x'}} \rho(x)\rho(x') \frac{\pi_{2|1}(x,x') - \pi_{2|1}(x)\pi_{2|1}(x')}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x')},
\end{aligned} \tag{4}$$

where j are identifiers of the sampling strata and

$$\rho(x) = \dot{e}^{(D)}(x) + \mathbf{\Delta_{Z^{(1)}}}' \tilde{\mathbf{G}}_{\beta_+^{(1)}}(x) \dot{e}^{(D_+)}(x) \tag{5}$$

- $\tilde{\mathbf{G}}_{\beta_+^{(1)}}(x) = \left[\mathbf{Z}_+^{(\mathbf{1})} \mathbf{\Sigma}_+ \mathbf{\Pi}_+ \mathbf{Z}_+^{(\mathbf{1})\prime}\right]^{-1} \frac{\mathbf{Z}_+^{(\mathbf{1})}(x)}{\pi^*(x)\sigma^2(x)}$,
- $\dot{e}^{(D)}(x) = \frac{\sum_{k=1}^{m} I_{kD} \dot{e}_k}{m}$ is cluster-level residual calculated with plot-level residuals set to 0 for plots outside the estimation cell D,
- $\dot{e}_k = y_k - \mathbf{Z}_{\mathbf{k}}^{(\mathbf{1})}(x)' \tilde{\mathbf{G}}_{\beta_+^{(1)}} \mathbf{\Pi}_+ \mathbf{Y}_+'$ is the corresponding plot-level residual,
- $\dot{e}^{(D_+)}(x) = \frac{\sum_{k=1}^{m} I_{kD_+} \dot{e}_k}{m}$ is the cluster-level residual calculated with plot-level residuals set to 0 is outside the parameterisation region D_+.

Note that the first two terms of (4) reflect the part of the variance due to the first phase of sampling (in geographic space), so they correspond to the Horvitz-Thompson theorem for continuous populations [Cordy, 1993]. The third and fourth terms are due to the second phase of sampling from the finite set of the first phase points, so they match the finite Horvitz-Thompson theorem, see the π-estimator by Särndal *et al.* [2003, pp. 42-48]. The derivation of variance can be found in Annex A.2.

The term $\pi^*(x, x')$ is the pairwise density for simultaneous selection of two distinct points x and x' considering both phases of sampling. It is the product of the pairwise density $\pi_1(x, x')$ of selecting the same points in the first phase (sampling the infinite population of points) and the conditional pairwise probability $\pi_{2|1}(x, x')$ of selecting again these two points in the second phase (sampling the finite population of points chosen in the first phase).

Note that x and x' always come from the same stratum. The pairs between strata are excluded by Eq. (4) to speed up the computation. They do not add to the variance estimator, because pairwise densities equal the product of single densities if both x and x' come from different strata. The contribution of these pairs to the second and fourth terms of (4) would be zero because the subtraction of the product of single densities from the pairwise density gives a zero value.

The pairwise density of the first phase is defined by

$$\pi_1(x, x') = \sum_{k=1}^{n_{1,j}} \sum_{\substack{l=1 \\ l \neq k}}^{n_{1,j}} f^{(j)}_{1,k,l}(x, x') \tag{6}$$

where $f^{(j)}_{1,k,l}(x, x')$ is the joint probability density for the selection of points x and x' as the k-th and l-th points of the sample in stratum j, see [Cordy, 1993]. It is convenient and usually valid to assume that the probability of selecting any point in the infinite population does not depend on the points that have been selected in the previous steps, so there is only one joint probability density $f^{(j)}_{1,k,l}(x, x')$, and we can write

$$\pi_1(x, x') = n_{1,j}(n_{1,j} - 1) f_{1,j}(x, x'). \tag{7}$$

Assuming the selection of x does not depend on the selection of any x' distinct from x, we get

$$\pi_1(x, x') = n_{1,j}(n_{1,j} - 1) f_{1,j}(x) f_{1,j}(x'), \tag{8}$$

where $f^{(j)}(x)$ and $f_j(x')$ are marginal probability densities of selecting x and x' respectively out from stratum j. From the definition of the inclusion density

$$\pi_1(x) = \sum_{k=1}^{n_1} f^{(j)}_{1,k}(x), \tag{9}$$

where $f^{(j)}_{1,k}(x)$ is the marginal probability density for the selection of point x in the k-th draw from the infinite population of points in the stratum j. If the chances of point selection do not depend on the order, we have just one marginal density $f^{(j)}_1(x)$ and

$$\pi_1(x) = n_{1,j} f_{1,j}(x). \tag{10}$$

Under all the above assumptions of independence and order irrelevance, the pairwise density is finally defined as

$$\pi_1(x, x') = \frac{(n_{1,j} - 1)}{n_{1,j}} \pi_1(x) \pi_1(x'). \tag{11}$$

This is a practical result because there is no need to specify $\pi_1(x, x')$ explicitly for all $\binom{n_{1,j}}{2}$ possible pairs of x and x' selected in stratum j. The specification of $\pi_1(x)$ for all sample points x is sufficient. Note that in the special case of sampling without replacement, the pairwise density is defined

$$f(x, x') = f(x) f(x'|x) = f(x') f(x|x'). \tag{12}$$

However, in fact $f(x'|x) = f(x')$ and $f(x|x') = f(x)$ because the measure of any point in the region is zero. So, equations (8), (11) are also valid for sampling without replacement.

With respect to the pairwise probability $\pi_{2|1}(x, x')$ it is also convenient to avoid explicit specification of the number of $\binom{n_{2,j}}{2}$ pairwise probabilities. This can be practically achieved if we simply consider that the second phase of sampling is uniform random without replacement. It leads to the formula

$$\pi_{2|1}(x, x') = \frac{n_{2,j}(n_{2,j} - 1)}{n_{1,j}(n_{1,j} - 1)}, \tag{13}$$

causing a potential overestimation of the variance. Finally, the $\pi^*(x, x')$ can be expressed using available sampling densities, the first- and the second-phase sample sizes

$$\pi^*(x, x') = \frac{n_{2,j}(n_{2,j} - 1)}{n_{1,j}^2}\pi_1(x)\pi_1(x') \tag{14}$$

$$= \frac{(n_{2,j} - 1)}{n_{2,j}}\pi^*(x)\pi^*(x'). \tag{15}$$

The estimators presented in this section and in the following sections are usable for any probability sampling from a geographical continuum. In case of stratified sampling, the variances are estimated by J strata and then summed, which brings a significant computational advantage. The pairs of sample points with each point from a different stratum do not contribute to the overall variance, so the computational burden of double summations is reduced. For typographical reasons, inclusion densities and probabilities are not marked by stratum but are always defined within strata, i.e., both x and x' come from one stratum.

2.2 Two-phase regression estimator of the total with the first-phase and exhaustive auxiliaries

This estimator uses two types of auxiliaries: $\mathbf{Z}_+^{(\mathbf{0})}(x)$ available at any point $x \in D_+$ (p_0 exhaustive auxiliaries extracted from a map), and $\mathbf{Z}_+^{(\mathbf{1})}(x)$ available only at points $x \in s_1$ (p_1 auxiliaries obtained in the first phase of sampling), but not everywhere in $D_+ \supset D$. In total, there are $p = p_0 + p_1$ auxiliaries involved in this estimator. It is assumed that, compared to exhaustive auxiliaries, the acquisition of $\mathbf{Z}_+^{(\mathbf{1})}(x)$ is more expensive, but is compensated for by higher correlations with field data leading to adequate precision gains. The estimator is calculated according to the equation:

$$\hat{t}_{y,2pm} = \mathbf{I_D \Pi_D Y'_D} + \boldsymbol{\Delta}_{\mathbf{z}}{}'\tilde{\mathbf{G}}_{\beta_+}\boldsymbol{\Pi}_+\mathbf{Y}'_+ + \boldsymbol{\Delta}'_{\mathbf{z^{(0)}}}\tilde{\mathbf{G}}_{\beta_+^{(\mathbf{0})}}\boldsymbol{\Pi}_+\mathbf{Y}'_+ \tag{16}$$

where

- $\boldsymbol{\Delta}_{\mathbf{z^{(0)}}} = \mathbf{t}_{\mathbf{z^{(0)}}} - \hat{\mathbf{t}}_{\mathbf{s_1 z^{(0)}}} = \mathbf{t}_{\mathbf{z^{(0)}}} - \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{0})}(x)}{\pi_1(x)}$ is the $p_0 \times 1$ difference between the true total of the vector of exhaustive auxiliaries and its single-phase estimate using the 1st-phase sample s_1,
- $\mathbf{Z}_{\mathbf{D}}^{(\mathbf{0})}(x) = \frac{\sum_{k=1}^{m} I_{kD}\mathbf{Z}_{\mathbf{k}}^{(\mathbf{0})}}{m}$ where $\mathbf{Z}_{\mathbf{k}}^{(\mathbf{0})}$ is $p_0 \times 1$ vector of exhaustive auxiliaries at plot k,
- $\boldsymbol{\Delta}_{\mathbf{z}} = \hat{\mathbf{t}}_{\mathbf{s_1,z}} - \hat{\mathbf{t}}_{\mathbf{s_2,z}} = \sum_{x \in s_1} \frac{\mathbf{Z_D}(x)}{\pi_1(x)} - \sum_{x \in s_2} \frac{\mathbf{Z}_D(x)}{\pi^*(x)}$ is the $p \times 1$ vector of differences of the single-phase estimates of total of the full auxiliary vector $\mathbf{Z_D}(x)$ between the first-phase and second-phase samples,

- $\mathbf{Z_D}(x) = \frac{\sum_{k=1}^{m} I_{kD}\mathbf{Z}_k}{m}$ with m corresponding to the nominal number of plots within a cluster and $\mathbf{Z}_k$ is the full auxiliary vector at plot k, it is an aggregation of the exhaustive and 1st-phase auxiliaries $\mathbf{Z_k} = \left\{\mathbf{Z_k}^{(0)}, \mathbf{Z_k}^{(1)}\right\}$, $\mathbf{Z_D}(x) = \left\{\mathbf{Z_D}^{(0)}(x), \mathbf{Z_D}^{(1)}(x)\right\}$,
- $\tilde{\mathbf{G}}_{\beta_+^{(\mathbf{0})}}$ is the $p_0 \times n_2$ matrix defined by

$$\tilde{\mathbf{G}}_{\beta_+^{(\mathbf{0})}} = \left[\mathbf{Z}_+^{(\mathbf{0})}\mathbf{\Sigma}_+\mathbf{\Pi}_+\mathbf{Z}_+^{(\mathbf{0})\prime}\right]^{-1}\mathbf{Z}_+^{(\mathbf{0})}\mathbf{\Sigma}_+, \tag{17}$$

- $\mathbf{Z}_+^{(0)}$ is the $p_0 \times n_2$ matrix of exhaustive auxiliaries available everywhere in $D_+ \supset D$, the columns of $\mathbf{Z}_+^{(\mathbf{0})}$ correspond to $p_0 \times 1$ auxiliary vector $\mathbf{Z}_+^{(\mathbf{0})}(x)$ at the sample point x,
- $\mathbf{Z}_+^{(0)}(x) = \frac{\sum_{k=1}^{m} I_{kD_+}\mathbf{Z_k^{(0)}}}{m}$, I_{kD_+} takes value 1 if and only if the centre of plot k belongs to the parameterisation region D_+,
- the $\tilde{\mathbf{G}}_{\beta_+}$ is the $p \times n_2$ matrix defined by

$$\tilde{\mathbf{G}}_{\beta_+} = \left[\mathbf{Z}_+\mathbf{\Sigma}_+\mathbf{\Pi}_+\mathbf{Z}_+'\right]^{-1}\mathbf{Z}_+\mathbf{\Sigma}_+, \tag{18}$$

- and the $\mathbf{Z}_+ = \left\{\mathbf{Z}_+^{(\mathbf{0})},\ \mathbf{Z}_+^{(\mathbf{1})}\right\}$ is the $p \times n_2$ full vector of auxiliaries.

The variance of $\hat{t}_{y,2pm}$ is calculated by:

$$\begin{aligned}
\hat{\mathbb{V}}\left(\hat{t}_{y,2pm}\right) = &\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\frac{\phi(x)^2}{\pi^*(x)\pi_1(x)} + \\
&+\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\phi(x)\phi(x')\frac{\pi_1(x,x') - \pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} + \\
&+\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\left[\frac{\omega(x)}{\pi^*(x)}\right]^2\left[1-\pi_{2|1}(x)\right] + \\
&+\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\omega(x)\omega(x')\frac{\pi_{2|1}(x,x') - \pi_{2|1}(x)\pi_{2|1}(x')}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x')},
\end{aligned} \tag{19}$$

where

$$\phi(x) = e^{(D)}(x) + \mathbf{\Delta_{z^{(0)}}}'\tilde{\mathbf{G}}_{\beta_+^{(0)}}(x)e^{(D_+)}(x) \tag{20}$$

and

- $\tilde{\mathbf{G}}_{\beta_+^{(0)}}(x) = \left[\mathbf{Z}_+^{(\mathbf{0})}\mathbf{\Sigma}_+\mathbf{\Pi}_+\mathbf{Z}_+^{(\mathbf{0})\prime}\right]^{-1}\frac{\mathbf{Z}_+^{(\mathbf{0})}(x)}{\pi^*(x)\sigma^2(x)}$,
- $e^{(D)}(x) = \frac{\sum_{k=1}^{m} I_{kD}e_k}{m}$ is the cluster-level residual using plot residuals of a model with exhaustive auxiliaries that are zeroed if the particular plot is outside D,
- $e_k = y_k - \mathbf{Z_k^{(0)}}'(x)\tilde{\mathbf{G}}_{\beta_+}^{(\mathbf{0})}\Pi_+Y_+'$ is the corresponding plot-level residual,

- $e^{(D_+)}(x) = \frac{\sum_{k=1}^{m} I_{kD_+} e_k}{m}$ is the cluster-level residual using plot residuals set to 0 outside of parameterisation region D_+,

$$\omega(x) = \ddot{e}^{(D)}(x) + \mathbf{\Delta_z} \tilde{\mathbf{G}}_{\beta_+}(x) \ddot{e}^{(D_+)}(x) \tag{21}$$

- $\tilde{\mathbf{G}}_{\beta_+}(x) = \left[\mathbf{Z}_+ \mathbf{\Sigma}_+ \mathbf{\Pi}_+ \mathbf{Z}'_+\right]^{-1} \frac{\mathbf{Z}_+(x)}{\pi^*(x)\sigma^2(x)}$,
- $\ddot{e}^{(D)}(x) = \frac{\sum_{k=1}^{m} I_{kD} \ddot{e}_k}{m}$ is the cluster-level residual using plot residuals of a model using both auxiliary types (full vector of auxiliaries) that are zeroed if the particular plot is outside D,
- $\ddot{e}_k = y_k - \mathbf{Z}'_{\mathbf{k}}(x) \tilde{\mathbf{G}}_{\beta_+} \mathbf{\Pi}_+ \mathbf{Y}'_+$ is the corresponding plot-level residual,
- $\ddot{e}^{(D_+)}(x) = \frac{\sum_{k=1}^{m} I_{kD_+} \ddot{e}_k}{m}$ is the cluster-level residual using plot residuals set to 0 outside of parameterisation region D_+,

The derivation of the variance estimator is analogous to the variance of the two-phase total estimator without exhaustive auxiliaries presented in Annex A.2.

Although there is no guarantee, the estimator (16) can lead to additional precision gains compared to (2). In the typical case, a forest mask can be used as an exhaustive auxiliary. A map of the parameterisation region D_+ can also be used with the vector $\mathbf{Z}_+^{(0)}$ constructed by using the indicator variable $I_{kD_+} = 1 \iff x \in D_+$ which balances[1] the estimate towards the known total area of D_+.

[1] If the single-phase estimator overestimates the area of D_+ (there is an upward imbalance in the sample) the regression estimator will correct the single phase estimator of y downwards, and vice versa.

3 Estimation of ratios

For almost every total variable, it is possible to define a meaningful ratio counterpart. For example, the total Above-Ground Biomass (AGB) has a related target variable AGB per hectare of forest area. The rate of change of a parameter can also be estimated as a ratio. The ratios are defined as follows:

$$r_{1,2} = \frac{t_{y_1}}{t_{y_2}}, \tag{22}$$

where in our example t_1 can be the total AGB and t_2 can be the total forest area. The estimator of a ratio is calculated according to the equation

$$\hat{r}_{1,2} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}}. \tag{23}$$

The estimators in the numerator and the denominator of (23) can be of different types, and they can use different samples, with or without any intersection.

With respect to variance derivation, the ratio estimator can be approximated by using the first-order Taylor series

$$\hat{r}_{1,2} \approx r_{1,2} + \frac{1}{t_2}(\hat{t}_{y_1} - t_{y_1}) - \frac{r_{1,2}}{t_{y_2}}(\hat{t}_{y_2} - t_{y_2}), \tag{24}$$

so, for the approximate variance estimator, one obtains

$$\hat{\mathbb{V}}\left(\hat{r}_{1,2}\right) = \frac{1}{\hat{t}_2^2}\hat{\mathbb{V}}\left(\hat{t}_{y_1} - \hat{r}_{1,2}\hat{t}_{y_2}\right), \tag{25}$$

If the numerator and denominator use the same samples, variance estimators can be derived from (25) for a specific combination of estimator types in the numerator and denominator[2]. Otherwise, the variance needs to be evaluated by pieces according to the formula

$$\hat{\mathbb{V}}\left(\hat{r}_{1,2}\right) = \frac{1}{\hat{t}_2^2}\hat{\mathbb{V}}(\hat{t}_{y_1}) + \frac{\hat{r}_{1,2}^2}{\hat{t}_2^2}\hat{\mathbb{V}}(\hat{t}_{y_2}) - \frac{\hat{r}_{1,2}}{\hat{t}_2^2}\hat{\mathbb{C}}(\hat{t}_{y_1}, \hat{t}_{y_2}). \tag{26}$$

Both variances as well as the covariance depend on the estimator type in the numerator and the denominator of the ratio. The following sections provide solutions for combinations of different estimator types and using identical or different samples in the numerator and denominator.

3.1 Ratio of single-phase estimators using an identical sample

If the numerator and denominator are single-phase estimators (with or without exhaustive auxiliaries) and the corresponding samples are identical $s^{(1)} = s^{(2)} = s$ the variance estimation follows earlier derivations by Adolt *et al.* [2018].

Equation (25) is applied stratum-wise

$$\hat{\mathbb{V}}\left(\hat{r}_{1,2}\right) = \frac{1}{\hat{t}_2^2}\sum_{j=1}^{J}\hat{\mathbb{V}}\left(\hat{t}_{y_1,j} - \hat{r}_{1,2}\hat{t}_{y_2,j}\right), \tag{27}$$

[2]Estimators with or without certain type of auxiliaries, exhaustive or 1st-phase.

and the variance estimator is calculated

$$\begin{aligned}\hat{\mathbb{V}}(\hat{r}_{1,2}) = &\sum_{j=1}^{J}\sum_{x\in s_j}\left[\frac{z(x)}{\pi(x)}\right]^2 \\ &+\sum_{j=1}^{J}\sum_{x\in s_j}\sum_{\substack{x\in s_j\\ x\neq x'}} z(x)z(x')\frac{\pi(x,x')-\pi(x)\pi(x')}{\pi(x,x')(x)\pi(x')},\end{aligned} \tag{28}$$

where the term $\pi(x)$ is the inclusion density commonly used for the total estimators in the numerator and denominator of $\hat{r}_{1,2}$ and $\pi(x,x')$ is the pairwise inclusion density conveniently defined with the meaningful assumptions made in Section 2.1

$$\pi(x,x') = \frac{(n_j-1)}{n_j}\pi(x)\pi(x'). \tag{29}$$

The residual variable $z(x)$ is defined by the particular combination of the types of single-phase estimator in the numerator and denominator (with or without exhaustive auxiliaries), see table 1.

Table 1: *The definition of residual variable $z(x)$ depending on the combination of estimator type in the numerator and denominator of a single-phase ratio (with or without exhaustive auxiliaries)*

numerator	denominator	$z(x)$
1p	1p	$y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)$
1p	1pm	$y_1^{(D)}(x) - \hat{r}_{1,2}\phi_2^{(D)}(x)$
1pm	1p	$\phi_1(x) - \hat{r}_{1,2}y_2^{(D)}(x)$
1pm	1pm	$\phi_1(x) - \hat{r}_{1,2}\phi_2^{(D)}(x)$

The terms $\phi(x)$ are defined by the equation (20) on page 11.

3.2 Ratio of single-phase estimators using unequal samples

This section applies to ratios, defined by single-phase totals with or without the use of exhaustive auxiliaries. The estimators in the numerator and denominator use different samples $s^{(1)}$ and $s^{(2)}$, which may $(s^{(1)} \cap s^{(2)} \neq \emptyset)$ but do not necessarily have to intersect $(s^{(1)} \cap s^{(2)} = \emptyset)$.

The variance calculation follows (26) assuming J sampling strata, which leads to

$$\hat{\mathbb{V}}\left(\hat{r}_{1,2}\right) = \frac{1}{\hat{t}_{y_2}^2}\sum_{j=1}^{J}\hat{\mathbb{V}}(\hat{t}_{y_1,j}) + \frac{\hat{r}_{1,2}^2}{\hat{t}_{y_2}^2}\sum_{j=1}^{J}\hat{\mathbb{V}}(\hat{t}_{y_2,j}) - \frac{\hat{r}_{1,2}}{\hat{t}_{y_2}^2}\sum_{j=1}^{J}\hat{\mathbb{C}}(\hat{t}_{y_1,j},\hat{t}_{y_2,j}). \tag{30}$$

The variances within strata of the numerator and denominator are evaluated according to the estimator type by the following equation

$$\begin{aligned}\hat{\mathbb{V}}(\hat{t}_{y,j}) = &\sum_{x\in s_j}\left[\frac{\tau(x)}{\pi(x)}\right]^2 \\ &+\sum_{x\in s_j}\sum_{\substack{x\in s_j\\ x\neq x'}} \tau(x)\tau(x')\frac{\pi(x,x')-\pi(x)\pi(x')}{\pi(x,x')(x)\pi(x')},\end{aligned} \tag{31}$$

which is the HTC by Cordy [1993] applied by sampling stratum. If no exhaustive auxiliaries are used, then the term $\tau(x) = y^{(D)}(x)$, it corresponds to the value of attribute density observed on point x. If the modified direct generalised regression estimator by Adolt *et al.* [2018] is applied, we set $\tau(x) = \phi(x)$, where $\phi(x)$ is defined by (20) on page 11.

The covariances within the last term of (30) for a given stratum can be estimated by

$$\begin{aligned}\hat{\mathbb{C}}\left[\hat{t}_{y_{1,j}}, \hat{t}_{y_{2,j}}\right] &= \sum_{x \in s^{(m)}} \frac{\tau_1(x)\tau_2(x)}{\pi^{(1)}(x)\pi^{(2)}(x)} \\ &= \sum_{x \in s_j^{(1)}} \sum_{\substack{x' \in s_j^{(2)} \\ x' \neq x}} \frac{\tau_1(x)\tau_2(x')}{\pi^{(1)}(x)\pi^{(2)}(x')} \times \frac{\pi^{(1,2)}(x,x') - \pi^{(1)}(x)\pi^{(2)}(x')}{\pi^{(1,2)}(x,x')}. \end{aligned} \tag{32}$$

The terms $\tau_1(x)$ and $\tau_2(x)$ are defined according to table 2, depending on whether exhaustive auxiliaries are used in the nominator, in the denominator, or in both total estimators defining the ratio.

Table 2: *The $\tau_1(x)$ and $\tau_2(x)$ terms in covariance formula of single-phase totals with or without exhaustive auxiliaries*

combination of estimator types		$\tau_1(x)$	$\tau_2(x)$
$\hat{t}_{y_1,j}$	$\hat{t}_{y_2,j}$		
1p	1p	$y_1^{(D)}(x)$	$y_2^{(D)}(x)$
1p	1pm	$y_1^{(D)}(x)$	$\phi_2(x)$
1pm	1p	$\phi_1(x)$	$y_2^{(D)}(x)$
1pm	1pm	$\phi_1(x)$	$\phi_2(x)$

The pairwise inclusion density $\pi^{(1,2)}(x,x')$ refers to the simultaneous inclusion of points x and x' in the corresponding samples $s_j^{(1)}$ and $s_j^{(2)}$. It is defined by equation

$$\pi^{(1,2)}(x,x') = \pi^{(1,2)}(x_i,x_l) = \sum_{x_i \in s_j^{(1)}} \sum_{\substack{x_l \in s_j^{(2)} \\ x_l \neq x_i}} f_{i,l}^{(j)}\left[x_i \in s_j^{(1)}, x_l \in s_j^{(2)}\right], \tag{33}$$

where $f_{i,l}^{(j)}\left[x_i \in s_j^{(1)}, x_l \in s_j^{(2)}\right]$ is the joint probability density of including point $x = x_i$ as the i-th element of $s_j^{(1)}$, and simultaneously including point $x' = x_l$ as the l-th element of $s_j^{(2)}$. In the context of forest monitoring, the joint probability density does not depend on the order of points x and x' in the samples. Thus, a single joint probability density $f^{(j)}\left[x \in s_j^{(1)}, x' \in s_j^{(2)}\right]$ can be used, and the pairwise inclusion density is defined

$$\pi^{(1,2)}(x,x') = (n_j^{(1)} n_j^{(2)} - n_m) f^{(j)}\left[x \in s_j^{(1)}, x' \in s_j^{(2)}\right], \tag{34}$$

where $n_j^{(1)}$ is the size of the sample $s_j^{(1)}$ in sampling stratum j, $n_j^{(2)}$ is the size of sample $s_j^{(2)}$ and n_m is the size of the matched sample $s_j^{(m)}$ defined later in this section.

In the monitoring system proposed by PathFinder, yearly (or even seasonal) samples (panels) are merged, so samples $s_j^{(1)}$ and $s_j^{(2)}$ corresponding to the desired periods of

estimation are obtained. One of the possible information needs is to estimate the relative change of a parameter between two consecutive periods that overlap, and so do the corresponding merged samples. This is precisely the situation where ratios of two totals using unequal sample are needed. Now, imagine the two samples $s_j^{(1)}$ and $s_j^{(2)}$ that are unions of several panels

$$s_j^{(1)} = \bigcup_{qj \in \mathcal{Q}_j} s_{qj} \qquad s_j^{(2)} = \bigcup_{vj \in \mathcal{V}_j} s_{vj} \tag{35}$$

The matched sample is defined as the union of all panels included in $s_j^{(1)}$, as well as $s_j^{(2)}$

$$s_j^{(m)} = \bigcup_{\substack{s_{hj} \in s_j^{(1)} \\ s_{hj} \in s_j^{(2)}}} s_{hj} \tag{36}$$

The derivation of simple inclusion densities for the merged samples is described in Section 6.4. The pairwise inclusion density for distinct points $x \in s_{qj} \subseteq s_j^{(1)}$ and $x' \in s_{qv} \subseteq s_j^{(2)}$ is derived based on the joint probability densities of the original samples s_{qj} and s_{qv} by the following equation

$$f^{(j)}\left[x \in s_j^{(1)}, x' \in s_j^{(2)}\right] = \begin{cases} s_{qj} \neq s_{qv} & \frac{1}{c_q} f_{qj}(x \in s_{qj}) \times \frac{1}{c_v} f_{vj}(x' \in s_{vj}) \\ s_{qj} = s_{vj} \wedge x \neq x' & \frac{1}{c_q}\frac{1}{c_v} f_{qj}(x \in s_{qj}, x' \in s_{qj}) \end{cases} \tag{37}$$

where c_q is the number of distinct panels belonging to $s_j(1)$, and c_v is the number of distinct panels belonging to $s_j^{(2)}$, and the indices q indicate the original sample from which point x comes from, and likewise for the index v and the point x'.

In the PathFinder platform, the joint probability density is approximated by the product of marginal densities, that is, we have $f_{qj}(x \in s_{qj}, x' \in s_{qj}) \approx f_{qj}(x \in s_{qj}) f_{qj}(x' \in s_{qj})$, as if the selection of the two points was independent. NFIs typically use a sort of spatially restricted sampling (for example, systematic or spatially stratified), and this approximation will likely cause an overestimation of covariance that will be to some extent compensated by overestimation of variances; see formula (26). With this approach, the pairwise density is calculated

$$\pi^{(1,2)}(x, x') = \frac{n_j^{(1)} n_j^{(2)} - n_j^{(m)}}{n_j^{(1)} n_j^{(2)}} \pi^{(1)}(x) \pi^{(2)}(x'). \tag{38}$$

Imagine that there is no shared part of the samples, so we have $n_j^{(m)} = 0$ and

$$\pi^{(1,2)}(x, x') = \pi^{(1)}(x) \pi^{(2)}(x). \tag{39}$$

This is an intuitive result for independent samples that leads to zero covariance according to (32).

Now, suppose that both totals use the same sample that is $n_j^{(1)} = n_j^{(2)} = n_j^{(m)}$ and the pairwise density

$$\pi^{(1,2)}(x, x') = \frac{n_j^{(m)} - 1}{n_j^{(m)}} \pi^{(1)}(x) \pi^{(2)}(x') \tag{40}$$

corresponds to its form (29) for matched samples.

3.3 Ratio of two-phase estimators using an identical sample

In this section, variance estimators are developed for a ratio of two-phase estimators (2) or (16) in the numerator and denominator.

The variance of any two-phase estimator $\hat{\theta}_{2p}$, be it total or ratio, is defined by the following equation:

$$\mathbb{V}(\hat{\theta}_{2p}) = \mathbb{V}_1\left[\mathbb{E}_{2|1}(\hat{\theta}_{2p})\right] + \mathbb{E}_1\left[\mathbb{V}_{2|1}(\hat{\theta}_{2p})\right]. \tag{41}$$

So, it is a sum of the variance of the conditional expectation given the first-phase sample $\mathbb{E}_{2|1}(\theta_{2p})$ and the expectation over the first phase of the second-phase variance given the first-phase sample $\mathbb{V}_{2|1}(\hat{\theta}_{2p})$, see Särndal et al. [2003, p. 136].

The estimator of variance of the ratio is given by the formula:

$$\hat{\mathbb{V}}\left(\hat{r}_{1,2}\right) = \frac{1}{\hat{t}_2^2}\left\{\begin{array}{l} \displaystyle\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\frac{\gamma_1^2(x)}{\pi^*(x)\pi_1(x)} + \\ \displaystyle\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\gamma_1(x)\gamma_1(x')\frac{\pi_1(x,x')-\pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} + \\ \displaystyle\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\left[\frac{\gamma_2(x)}{\pi^*(x)}\right]^2\left[1-\pi_{2|1}(x)\right] + \\ \displaystyle\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\gamma_2(x)\gamma_2(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x')}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x')} \end{array}\right\}. \tag{42}$$

The terms $\gamma_1(x)$ and $\gamma_2(x)$ are specific for the combination of estimator types in the numerator and denominator of $\hat{r}_{1,2}$, see the table 3 below.

Table 3: *The terms $\gamma_1(x)$ and $\gamma_2(x)$ according to (42) for the variance of two-phase ratio*

numerator	denominator	$\gamma_1(x)$	$\gamma_2(x)$	annex
2p	2p	$y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)$	$\rho_1(x) - \hat{r}_{1,2}\rho_2(x)$	B.1
2p	2pm	$y_1^{(D)}(x) - \hat{r}_{1,2}\phi_2(x)$	$\rho_1(x) - \hat{r}_{1,2}\omega_2(x)$	B.2
2pm	2p	$\phi_1(x) - \hat{r}_{1,2}y_2^{(D)}(x)$	$\omega_1(x) - \hat{r}_{1,2}\rho_2(x)$	B.3
2pm	2pm	$\phi_1(x) - \hat{r}_{1,2}\phi_2(x)$	$\omega_1(x) - \hat{r}_{1,2}\omega_2(x)$	B.4

The terms ρ, ϕ, and ω are defined by (5), (20) and (21) (pages 8, 11, and 12). Lower indexes are used to link to the respective term of the numerator (1) and the denominator (2). The derivation of $\gamma_1(x)$ and $\gamma_2(x)$ for each combination of estimator type in the numerator and denominator can be found in annex B.

The inclusion densities $\pi_1(x)$, $\pi^*(x)$, $\pi_1(x,x')$, $\pi^*(x,x')$, and inclusion probabilities $\pi_{2|1}(x)$ and $\pi_{2|1}(x,x')$ are defined in the same way as in Section 2.1.

3.4 Two-phase estimator in numerator, denominator or in both using unequal samples

The numerator of the ratio is estimated on the basis of samples $s_2^{(1)}$, and if it is a two phase estimator, also using $s_1^{(1)}$. The denominator uses samples $s_2^{(2)}$ and potentially also $s_1^{(2)}$. These samples and also the matched samples $s_1^{(m)}$, and $s_2^{(m)}$ are defined in analogy to (35), and (36).

The variance is estimated based on (30). The within-strata variances are estimated by (31) (single-phase estimators), or by (4) and (19) for two-phase estimators.

Annex C contains derivations of covariances of two-phase totals. The covariance of single-phase and two-phase estimators is a special case, because the former can be cast to the latter, by having all first-phase points selected in the second phase ($s_1 = s_2$, $n_1 = n_2$, $\pi_{2|1}(x) = 1$).

The within-stratum covariance estimator is given by the equation

$$\begin{aligned}\hat{\mathbb{C}}\left[\hat{t}_{y_1,2p,j}, \hat{t}_{y_2,2p,j}\right] = &\sum_{x \in s_2^{(m)}} \frac{\tau_1(x)\tau_2(x)}{\pi_{2|1}^{(m)}(x)\pi_1^{(1)}(x)\pi_1^{(2)}(x)} + \\ &+ \sum_{x \in s_2^{(1)}} \sum_{\substack{x' \in s_2^{(2)} \\ x' \neq x}} \frac{\tau_1(x)\tau_2(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')} \times \frac{\pi_1^{(1,2)}(x,x') - \pi_1^{(1)}(x)\pi_1^{(2)}(x')}{\pi^{*(1,2)}(x,x')} + \\ &+ \sum_{x \in s_2^{(m)}} \frac{\tau_3(x)\tau_4(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)} \left[1 - \pi_{2|1}^{(m)}\right] + \\ &+ \sum_{x \in s_2^{(1)}} \sum_{\substack{x' \in s_2^{(2)} \\ x' \neq x}} \frac{\tau_3(x)\tau_4(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')} \times \frac{\pi_{2|1}^{(1,2)}(x,x') - \pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')}{\pi_{2|1}^{(1,2)}(x,x')}.\end{aligned} \tag{43}$$

For simplicity, the $_j$ indices for strata are not used in this section, except for the symbols of the two estimators. The inclusion densities $\pi_1^{(1)}(x)$, $\pi_1^{(2)}(x)$, $\pi^{*(1)}(x)$, $\pi^{*(2)}(x)$ are analogous to their counterparts in Section 2.1. The conditional inclusion densities $\pi_{2|1}^{(1)}(x)$, $\pi_{2|1}^{(1)}(x')$ and $\pi_{2|1}^{(1,2)}(x,x')$ and the pairwise densities $\pi_1^{(1,2)}(x,x')$ and $\pi^{*(1,2)}(x,x')$ are defined in Annex C.1.

In analogy to (38), the pairwise density $\pi_1^{(1,2)}(x,x')$ is calculated using the following equation

$$\pi_1^{(1,2)}(x,x') = \frac{n_1^{(1)} n_1^{(2)} - n_1^{(m)}}{n_1^{(1)} n_1^{(2)}} \pi_1^{(1)}(x)\pi_1^{(2)}(x'), \tag{44}$$

where $n_1^{(1)}$, $n_1^{(2)}$ and $n_1^{(m)}$ are the sizes of samples $s_1^{(1)}$, $s_1^{(2)}$ and $s_1^{(m)}$.

The terms $\tau_1(x)$, $\tau_2(x)$, $\tau_3(x)$ and $\tau_4(x)$ depend on the types of total estimators. Table 4 shows all possible combinations and the corresponding terms $\tau_1(x)$, $\tau_2(x)$, $\tau_3(x)$ and $\tau_4(x)$ within strata. The terms $\rho(x)$, $\phi(x)$, and $\omega(x)$ are defined by (5), (20) and (21) (pages 8, 11, and 12). The lower indices $_1$ and $_2$ refer to the total estimators $\hat{t}_{y_1,j}$ and $\hat{t}_{y_2,j}$, respectively.

Table 4: *The $\tau(x)$ for the calculation of covariance of total estimators, depending on the combination of estimator types*

combination of estimator types		$\tau_1(x)$	$\tau_2(x)$	$\tau_3(x)$	$\tau_4(x)$
$\hat{t}_{y_1,j}$	$\hat{t}_{y_2,j}$				
1p	2p	$y_1^{(D)}(x)$	$y_2^{(D)}(x)$	$y_1^{(D)}(x)$	$\rho_2(x)$
1p	2pm	$y_1^{(D)}(x)$	$\phi_2^{(D)}(x)$	$y_1^{(D)}(x)$	$\omega_2(x)$
1pm	2p	$\phi_1(x)$	$y_2^{(D)}(x)$	$y_1^{(D)}(x)$	$\rho_2(x)$
1pm	2pm	$\phi_1(x)$	$\phi_2^{(D)}(x)$	$y_1^{(D)}(x)$	$\omega_2(x)$
2p	1p	$y_1^{(D)}(x)$	$y_2^{(D)}(x)$	$\rho_1(x)$	$y_2^{(D)}(x)$
2p	1pm	$y_1^{(D)}(x)$	$\phi_2^{(D)}(x)$	$\rho_1(x)$	$y_2^{(D)}(x)$
2p	2p	$y_1^{(D)}(x)$	$y_2^{(D)}(x)$	$\rho_1(x)$	$\rho_2(x)$
2p	2pm	$y_1^{(D)}(x)$	$\phi_2^{(D)}(x)$	$\rho_1(x)$	$\omega_2(x)$
2pm	1p	$\phi_1(x)$	$y_2^{(D)}(x)$	$\omega_1(x)$	$y_2^{(D)}(x)$
2pm	1pm	$\phi_1(x)$	$\phi_2(x)$	$\omega_1(x)$	$y_2^{(D)}(x)$
2pm	2p	$\phi_1(x)$	$y_2^{(D)}(x)$	$\omega_1(x)$	$\rho_2(x)$
2pm	2pm	$\phi_1(x)$	$\phi_2(x)$	$\omega_1(x)$	$\omega_2(x)$

4 Differences of total estimators

Many information needs are related to the evaluation of the change in a target parameter. Often, there is a wish to estimate changes for periods (one year, for example) that are much shorter than the typical length of an NFI cycle (usually 5 to 10 years). Therefore, direct change estimators [McRoberts *et al.*, 2015] cannot be used despite the use of periodically surveyed plots. Another practical situation is the calculation of a difference between two moving-window estimates that use five yearly panels, but mutually shifted by one year. In this case, there are four panels shared between the two estimates and two panels, each one used for one estimate exclusively. A generic estimator of change capable of transition between direct (remeasured permanent plots), hybrid (mixture of remeasured and distinct sets of plots), and indirect (temporary and permanent plots without re-measurement in the period) is needed. This section deals with differences of totals. Solutions for ratios are presented in Section 5. The techniques can be used to estimate changes of a parameter in time, but also for mere differences between different parameters at the same or distinct time periods.

4.1 Difference of single-phase estimators of total

Assume the totals of two variables y_1 and y_2 and their totals over the study region D defined in analogy to (1). Their single-phase estimators without using exhaustive auxiliaries (a map) are defined:

$$\hat{t}_{y_1} = \frac{\sum_{x \in s^{(1)}} y_1^{(D)}(x)}{\pi^{(1)}(x)} \tag{45}$$

$$\hat{t}_{y_2} = \frac{\sum_{x \in s^{(2)}} y_2^{(D)}(x)}{\pi^{(2)}(x)} \tag{46}$$

Without any reduction in applicability, it can be assumed that y_1 and y_2 are the same variables, but at two time points. The difference of two totals t_{y_2} and t_{y_1} is estimated naturally by

$$\hat{\Delta}_{t_{y12}} = \hat{t}_{y_2} - \hat{t}_{y_1} \tag{47}$$

Suppose that $\hat{t}_{y_2}$ and $\hat{t}_{y_1}$ are calculated based on partially intersecting samples $s_2^{(1)}$ and $s_2^{(2)}$

$$\mathbb{V}(\hat{\Delta}_{t_{y12}}) = \mathbb{V}(\hat{t}_{y_1}) + \mathbb{V}(\hat{t}_{y_2}) - 2\mathbb{C}, (\hat{t}_{y_1}, \hat{t}_{y_2}) \tag{48}$$

and estimated by

$$\hat{\mathbb{V}}(\hat{\Delta}_{t_{y12}}) = \hat{\mathbb{V}}(\hat{t}_{y_1}) + \hat{\mathbb{V}}(\hat{t}_{y_2}) - 2\hat{\mathbb{C}}(\hat{t}_{y_1}, \hat{t}_{y_2}). \tag{49}$$

The covariance term plays a role in cases where the estimators are calculated based on overlapping (or identical) samples.

If the difference is over a region that spans the number of J sampling strata, the estimators $\hat{t}_{y_1}$, $\hat{t}_{y_2}$ and $\hat{\Delta}t_{y12}$ do not change. Naturally, the sampling densities must correspond to the stratum in which the given sample point x is located.

$$\hat{\mathbb{V}}(\hat{\Delta}_{t_{y12}}) = \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{t}_{y_{1,j}}) + \sum_{j=1}^{J} \hat{\mathbb{V}}(\hat{t}_{y_{2,j}}) - 2\sum_{j=1}^{J} \hat{\mathbb{C}}\left[\hat{t}_{y_{1,j}}, \hat{t}_{y_{2,j}}\right]. \tag{50}$$

For single-phase estimators (with or without auxiliaries) the within-strata variance in the first two terms of (50) are estimated by (31) in section 3.2. The within-strata covariances in the last term of (50) are estimated by (32).

4.2 Difference involving two-phase regression estimators of total

The variance of a difference involving at least one two-phase estimator (2) or (16) naturally follows (50). The first two terms are the estimates of variance, according to (4), (19) or (31) applied by sampling strata. The covariances in the last term must be estimated specifically for each combination of estimator types, see (43) in Section 3.4. Annex C derives the elements (τ terms) of the covariance of totals, involving two-phase.

5 Differences of ratio estimators

The estimator of a difference of two ratios is defined by the following equation:

$$\hat{\Delta}_{\hat{r}} = \hat{r}_{3,4} - \hat{r}_{1,2} \tag{51}$$

$$\hat{r}_{1,2} = \frac{\hat{t}_{y_1}}{\hat{t}_{y_2}} \qquad \hat{r}_{3,4} = \frac{\hat{t}_{y_3}}{\hat{t}_{y_4}} \tag{52}$$

In analogy to (48) the variance of $\hat{\Delta}_{\hat{r}}$ is calculated by

$$\hat{\mathbb{V}}(\hat{\Delta}_{\hat{r}}) = \hat{\mathbb{V}}(\hat{r}_{1,2}) + \hat{\mathbb{V}}(\hat{r}_{3,4}) - 2\hat{\mathbb{C}}\left[\hat{r}_{1,2}, \hat{r}_{3,4}\right]. \tag{53}$$

If both, the numerator and the denominator are single-phase estimators with or without exhaustive auxiliaries, then the variances $\hat{\mathbb{V}}(\hat{r}_{1,2})$ and $\hat{\mathbb{V}}(\hat{r}_{3,4})$ will be calculated

- according to the formula (28) if the numerator and denominator use the same sample
- according to the formulas (30), (31) and (32) if numerator and denominator use unequal samples with or without overlap

If the numerator, the denominator or both correspond to any of the two-phase regression estimators presented in section 2, then the variances of the two ratios will be estimated

- by (42) if both numerator and denominator are two-phase estimators and use identical first-phase and second-phase samples
- by (30), (31) or (4) or (19), and by (43) in any other case (unequal samples)

If both ratios are estimated using a common sample, then the covariance in (53) will be evaluated according to section 5.1. If this is not the case, the covariance calculation follows section 5.2.

5.1 Difference of ratios using a common sample

If an identical same sample (the same set of sample plots or clusters) is used in the numerator and denominator of both ratios, then the covariance term in (53) will be greatly simplified. If both ratios are composed of single-phase estimators of total, then the covariance term follows

$$\begin{aligned}\hat{\mathbb{C}}\left[\hat{r}_{1,2},\hat{r}_{3,4}\right] = &\sum_{j=1}^{(J)}\sum_{x\in s_{2,j}}\frac{z_{1,2}(x)z_{3,4}(x)}{\pi^2(x)} \\ &+\sum_{j=1}^{J}\sum_{x\in s_{2,j}}\sum_{\substack{x\in s_{2,j}\\ x\neq x'}} z_{1,2}(x)z_{3,4}(x')\frac{\pi(x,x')-\pi(x)\pi(x')}{\pi(x,x')\pi(x)\pi(x')},\end{aligned} \tag{54}$$

where the $z_{1,2}(x)$ and $z_{3,4}(x)$ correspond to residual variables defined according to the combination of the types of total estimators in the respective ratio, see table 1.

If both ratios use a two-phase estimator of total their numerator and denominator, then the covariance can be estimated by

$$\hat{\mathbb{C}}\left[\hat{r}_{1,2},\hat{r}_{3,4}\right] = \frac{\sum_{j=1}^{J}}{\hat{t}_{y_2}\hat{t}_{y_4}}\left\{\begin{array}{l}\displaystyle\sum_{x\in s_{2,j}}\frac{\gamma_1^{(1,2)}(x)\gamma_1^{(3,4)}(x)}{\pi^*(x)\pi_1(x)} + \\ \displaystyle\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\gamma_1^{(1,2)}(x)\gamma_1^{(3,4)}(x')\frac{\pi_1(x,x')-\pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} + \\ \displaystyle\sum_{x\in s_{2,j}}\frac{\gamma_2^{(1,2)}(x)\gamma_2^{(3,4)}(x)}{\left[\pi^*(x)\right]^2}\left[1-\pi_{2|1}(x)\right] + \\ \displaystyle\sum_{x\in s_{2,j}}\sum_{\substack{x'\in s_{2,j}\\ x'\neq x'}}\gamma_2^{(1,2)}(x)\gamma_2^{(3,4)}(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x')}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x')}\end{array}\right\} \tag{55}$$

where $\gamma(x)$ terms depend on the combination of estimator type in the numerator and the denominator of the respective ratio marked by $^{(1,2)}$ or $^{(3,4)}$ in the upper index. Table 3 shows the $\gamma_1(x)$ and $\gamma_2(x)$ for each combination. Note that as soon as both ratios do not use the same and the only sample, for their numerators as well as denominators the generic approach of the next section must be used. This also includes mixtures of single-phase and two-phase totals within or between the two ratios.

5.2 Generic difference of two ratios

If the two ratios are not estimated on the basis of the same sample, the covariance of the ratios must be approximated. This is also the case if the totals in the numerator and denominator of any of the two ratios use unequal samples. A solution is proposed based on the first-order Taylor linearisation of the covariance

$$\hat{\mathbb{C}}\left[\hat{r}_{1,2},\hat{r}_{3,4}\right]=\frac{1}{\hat{t}_{y_2}\hat{t}_{y_4}}\left\{\begin{array}{l}\hat{\mathbb{C}}\left[\hat{t}_{y_1},\hat{t}_{y_3}\right]-\hat{r}_{3,4}\hat{\mathbb{C}}\left[\hat{t}_{y_1},\hat{t}_{y_4}\right]- \\ -\hat{r}_{1,2}\hat{\mathbb{C}}\left[\hat{t}_{y_2},\hat{t}_{y_3}\right]+\hat{r}_{1,2}\hat{r}_{3,4}\hat{\mathbb{C}}\left[\hat{t}_{y_2},\hat{t}_{y_4}\right]\end{array}\right\} \tag{56}$$

So, a series of terms corresponding to covariance of the totals involved in the estimation of any of the two ratios is obtained. Assuming the existence of J sampling strata, the covariance is estimated by

$$\hat{\mathbb{C}}\left[\hat{r}_{1,2},\hat{r}_{3,4}\right]=\frac{2}{\hat{t}_{y_2}\hat{t}_{y_4}}\sum_{j=1}^{J}\left\{\begin{array}{l}\hat{\mathbb{C}}\left[\hat{t}_{y_{1,j}},\hat{t}_{y_{3,j}}\right]-\hat{r}_{3,4}\hat{\mathbb{C}}\left[\hat{t}_{y_{1,j}},\hat{t}_{y_{4,j}}\right]- \\ -\hat{r}_{1,2}\hat{\mathbb{C}}\left[\hat{t}_{y_{2,j}},\hat{t}_{y_{3,j}}\right]+\hat{r}_{1,2}\hat{r}_{3,4}\hat{\mathbb{C}}\left[\hat{t}_{y_{2,j}},\hat{t}_{y_{4,j}}\right]\end{array}\right\}. \tag{57}$$

The covariances of the total estimators within the last term are estimated by (32) for the single-phase estimators and by (43) if any of the two totals is a two-phase estimator. These two formulas also work if identical samples are used in the pair of totals. There are 16 possible combinations of single- or two-phase regression totals (with or without exhaustive auxiliaries) for each of the two ratios $\hat{r}_{1,2}$ and $\hat{r}_{3,4}$. This generates a total of $16^2 = 256$ difference estimators and their variances.

6 Implementation

6.1 Single-phase regression estimator of the total using auxiliaries in the form of maps

The modified direct generalised regression estimator using exhaustive auxiliaries in the form of maps [Adolt *et al.*, 2018] was implemented in the PostgreSQL extension *nfiesta_pg* (https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/home) since the finalisation of the EU Horizon 2020 Diabolo project (Distributed, integrated and harmonised forest information for bioeconomy outlooks, H2020-ISIB-2014-2, http://diabolo-project.eu). However, the practical calculation of this estimator was not convenient due to the complexities of field and auxiliary data, all the more so with the increasing number of combinations of auxiliaries and working models. Therefore, a graphic user interface has been implemented within PathFinder to facilitate the process of estimator configuration (assignment of data and working model to each desired estimate depending on region and period). The use case can be found under the following link: https://gitlab.com/nfiesta/nfiesta_gui/-/wikis/Use case for configuration of GREG-map totals and ratios

6.2 Generic estimator of a difference

Many information needs are related to the evaluation of change of a target parameter. Often, there is a wish to estimate changes for short periods (one year, for example) that are shorter than the typical length of an NFI cycle. Therefore, direct estimators of change cannot be used despite the use of periodically surveyed sample plots. However, there are situations where a mixture of permanent and temporal sample plots can be used for the change estimation. This situation can ideally be addressed by a generic estimator of change capable of transition between direct (remeasured permanent plots), hybrid (mixture of remeasured and distinct sets of plots) and indirect (temporary or permanent plots without remeasurement in the period). Differences of totals as well as ratios need to be implemented in all variants developed in this report. The current schema for the implementation logic can be found here: https://gitlab.com/nfiesta/nfiesta_gui/-/wikis/Use case for configuration of difference of two estimators. The generic estimator of difference has not been implemented before this report compilation.

6.3 Estimation for arbitrary geographical regions

The problem of NFI plot coordinates disclosure is a well-known issue that any integrated monitoring system must handle. Disclosure of NFI plot coordinates can lead to the violation of representativeness of the NFI samples breaking the design-based principles [Schadauer *et al.*, 2024]. A solution proposed by PathFinder is based on the assignment of inventory plots to the cells of the 1 km INSPIRE (Infrastructure for Spatial Information in the European Community) grid [INSPIRE Maintenance and Implementation Group (MIG), 2023] according to the position of the plot centres inside the 1 km cells. If the geographical regions for estimation are approximated by INSPIRE cells of 1 km (based on the position of the centroid of the cell inside the polygon of the geographical region), then the set of plots in these regions can be easily determined.

The covered area is the union of all geographical regions for which estimates can be produced. It is defined as the union of the geometries of all NUTS 0 (countries by Eurostat) that provided the NFI data. The union is approximated by 1 km INSPIRE

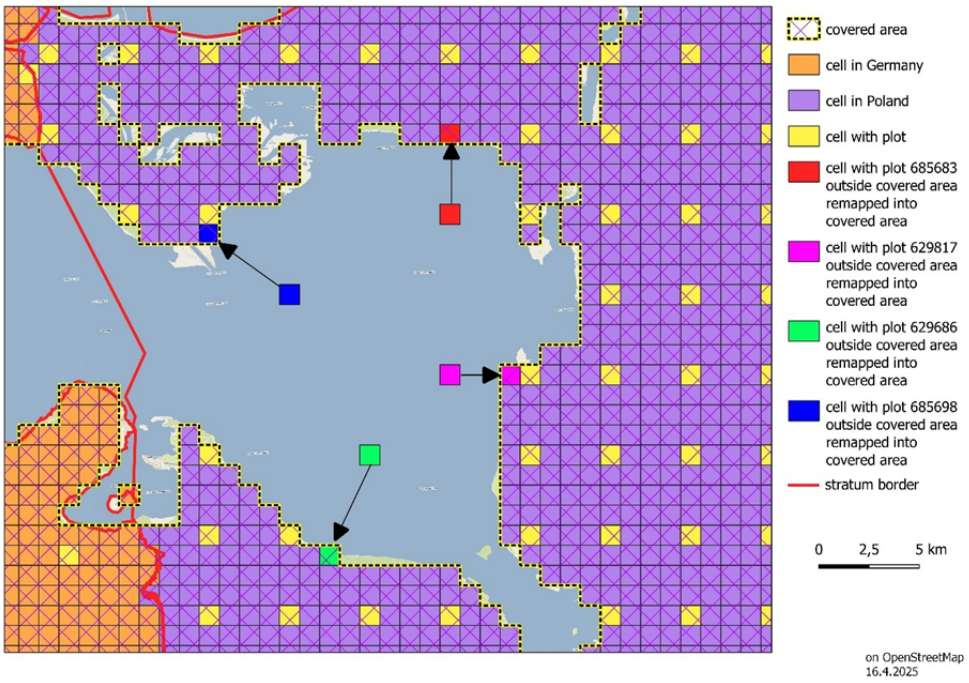

Figure 1: *Assignment of plots found in 1 km cells outside the covered area (approximation of NUTS0 by 1 km cells) to the nearest 1 km cells inside the covered area. The pale blue area in the centre of the picture corresponds to sea, so it is not part of the covered area. Courtesy of Vendula Hejlová, Czech Forestry Institue*

cells. Only cells with their centroid inside any of the NUTS0 geometries are preserved in the covered area. However, despite their position inside the covered area before its approximation, some plots may fall out of the estimation because they belong to a 1 km cell that is not part of the covered area (the centre of the cell does not fall into it). The solution is based on the reassignment of each such plot to the closest 1 km INSPIRE cell included in the approximated region of the country to which the plot belongs. In case of more candidate cells with the same distance, the one with lowest azimuth from the centroid of the original cell is chosen; see Figure 1.

In this way, the estimation for a country approximated by 1 km INSPIRE cells uses the full dataset of the country, and there is no underestimation due to omission of plots on the periphery of covered area. It should also be understood that the covered area does not necessarily correspond to the union of sampling strata provided by NFIs. This may be due to a too generous definition of the sampling strata that extend beyond the actual land territory of the country. For example, some countries include parts of their territorial waters to catch some small islets. Another common reason are mismatches caused by the different cartographic projections used for the definition of strata of individual countries. If the strata are then reprojected into a common coordinate system (used for the definition of the covered area), slight overlaps or gaps sometimes occur. Similar differences arise from various levels of generalisation of sampling strata (done by countries) and country geometries (sourced from Eurostat).

Because all estimation cells are approximated by the 1 km inspire grids, the estimates produced for a country often exclude some of the country's plots that fell into a 1 km cell assigned to a neighbouring country. And vice versa, some plots of the neighbouring country enter the estimation. Both effects compensate for each other, but they still cause

slight deviations from estimates produced using the country's full and the only dataset. This is of little relevance, because the role of an information system integrating NFI data of countries is not to produce national estimates, but estimates for arbitrary geographical regions going across country borders, and estimates for the whole continent.

6.4 Estimation for arbitrary time periods, temporal panel merging

Estimation for regions that cross country borders is challenging not only due to the mixture of various NFI designs, but also in terms of temporal synchronisation. Various countries use various time plans for their NFI surveys, that may even vary among several sampling strata. Therefore, it is advantageous to split the samples from each stratum into panels. Panels are spatially representative sets of single-plots or clusters that are surveyed according to a prescribed time plan. It is natural to use panels that were surveyed in the period for which estimates are required. Time is considered as one of the dimensions of sampling for which representativeness is to be achieved. Sampling weights are defined as the reciprocal values of sampling densities

$$w(x) = \pi^{-1}(x). \tag{58}$$

Assume $w_p(x)$ is the original sampling weight (in units of area) within the panel p. When merging panels, the original weights are divided by the m number of merged panels.

$$w(x) = \frac{w_p(x)}{m} \tag{59}$$

In addition, it is assumed that the period of each reference-year set (time between its beginning and end) is the same for all merged panels. This reweighting is an analogy to the calculation of separate estimates $\hat{\theta}_i$ for each panel followed by the calculation of the overall estimate as the arithmetic mean:

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta} i \tag{60}$$

This is principally the same approach as the moving average estimator suggested by Ene et al. [2018, p. 15], but with no need to calculate the partial estimates first.

The panel merging is done by stratum as part of the estimate configuration process, during which combinations of panels and particular reference year(s) are assigned to the each estimation task; see diagrams on the nFIESTA wiki pages:

- https://gitlab.com/nfiesta/nfiesta_gui/-/wikis/Use case for configuration of single-phase totals and ratios
- https://gitlab.com/nfiesta/nfiesta_gui/-/wikis/Use case for configuration of GREG-map totals and ratios

Addendum

A.1 Merging panels within the same reference period

In some situations, there might be more panels, say L panels, within the same stratum representing the same period of time. For example, there might be samples of permanent and temporal plots with different sample sizes, plot designs, etc. leading to different variances of estimators produced for each sample separately. Merging such panels in the way described in section 6.4 actually calculates the arithmetic average of the estimates. This is far from optimal, unless the variances of the individual estimates are practically the same. Often one would obtain an estimate that is worse than one or more of the uncombined estimates.

The solution is to combine the separate estimates by calculating the weighted average with weights inversely proportional to their variances

$$\hat{\theta} = \sum_{l=1}^{L} w_l \hat{\theta}_l, \tag{61}$$

where the weights w_l are defined

$$w_l = \frac{\hat{\mathbb{V}}[\hat{\theta}_l]^{-1}}{\sum_{l=1}^{L} \hat{\mathbb{V}}[\hat{\theta}_l]^{-1}}. \tag{62}$$

The resulting variance

$$\hat{\mathbb{V}}[\hat{\theta}] = \frac{1}{\sum_{l=1}^{L} \hat{\mathbb{V}}[\hat{\theta}_l]^{-1}} \tag{63}$$

is always lower than the variance of any of the single estimates, see Casella & Berger [2002, section 7.5]. This feature was not implemented in nFIESTA before the compilation of this report.

A.2 Model-assisted estimation with endogenous auxiliaries

Model-assisted estimators use auxiliary information to improve the precision of estimates. Generalized regression estimators with auxiliaries in the form of maps use a working model

$$y(x) = \mathbf{Z}'(x)\boldsymbol{\beta} + e(x) \tag{64}$$

where $y(x)$ is the target variable value at the sample point x, $\boldsymbol{\beta}$ is a vector $p \times 1$ of model parameters, $\mathbf{Z}(x)$ is a $p \times 1$ vector of auxiliaries (extracted from the map at the point x) and $e(x)$ is the error with normal distribution, zero mean value, and variance $\sigma^2(x)$. The variance may depend on the sample point, so it does not necessarily need to be constant. In practical applications, the model parameters are estimated from the sample data, so one gets

$$y(x) = \mathbf{Z}'(x)\hat{\boldsymbol{\beta}} + r(x) \tag{65}$$

where $\tilde{y}(x) = \mathbf{Z}'(x)\hat{\boldsymbol{\beta}}$ is the prediction of the working model and $r(x)$ is the residual. The model-assisted estimator can be expressed by the following equation

$$\hat{t}_{ma} = t_{\tilde{y}} + \hat{t}_r \tag{66}$$

where $t_{\tilde{y}}$ is the known total of predictions $\tilde{y}(x)$ and $\hat{t}_r$ is the single-phase estimate of the total of residuals $r(x)$. There is a non-zero variance of $\tilde{y}(x)$ and consequently $t_{\tilde{y}}$ because the estimates of model parameters vary from sample to sample (design-based variance). The vector of $\mathbf{Z}(x)$ is considered fixed (does not vary from sample to sample) thanks to the assumed independence of the (exogenous) auxiliaries and the sample field data. This is only the case if the sample data were not used for the creation of the map. However, it is not uncommon for maps to be produced using the same sample data as signatures (training set) for the map production. In such a case, the map is no more independent of the sample field data, and it is called endogenous. The vector of auxiliaries $\mathbf{Z}(x)$ is no longer fixed because its values depend on the sample. Now, the variances of $\tilde{y}(x)$ and $t_{\tilde{y}}$ are not zero because there is a variance of $\mathbf{Z}(x)$.

Breidt & Opsomer [2008] showed that the effect on the overall variance of the post-stratified estimator is negligible if classification is performed using generalised regression models fitted to the sample data. Maps are often created by more complex approaches such as machine learning, for which the effects are not easy to anticipate. A simulation study using semi-parametric (spline) and non-parametric (random forest) classifiers revealed that the post-stratified estimators worked also well in the endogenous setting as long as strata definitions are not optimised. If strata definitions are optimised variance underestimation has been observed [Tipton *et al.*, 2013]. Another study provided evidence that endogenous stratification works well for any non-parametric monotone model [Dahlke *et al.*, 2013]. These three studies support the practice of using sample data for the derivation of maps, which in turn are used as auxiliaries for model-assisted estimation.

A.3 The effects of over-fitting and model-assisted variance

Over-fitting means that a map is closer to reality in the training data positions (sample plots) than in the rest of the territory. An approach is suggested, so the model-assisted variance is not underestimated.

Assume that the sample plot i in position x_i is linked to a cell $\mathcal{C}_l$ in the study area $\mathcal{U}$. Note that $\forall l, l'$ such that $l \neq l'$ the relations $\mathcal{C}_l \cap \mathcal{C}_{l'} = \emptyset$ (cells do not intersect) and $\bigcup_{l=1}^{L} \mathcal{C}_l = \mathcal{U}$ (their union corresponds to the study area). There are up to $m(x)$ distinct cells to which any of the $m(x)$ plots of a cluster with anchor point x belongs. In the production of the map for a specific cell, no plots attached to the cell are used to train the classifier or derive the map values in any other way. If single plots are used, then one simply assumes clusters with one plot only ($\forall\, x,\ m\,(x) \ = \ 1$).

In the PathFinder platform, the cells $\mathcal{C}_l$ correspond to INSPIRE cells of 1 km by which one approximates any subregion of $\mathcal{U}$ where estimates must be calculated. Because only plot membership to 1 km INSPIRE cells is provided, all plots attached to the corresponding 1 km INSPIRE cell are left out of the training data. In this way, the impact of over-fitting causing underestimation of model-assisted variances can be avoided or at least reduced to a possible minimum.

References

Adolt, R., Fejfar, J., Lanz, A., & Ene, L., T. 2018. *nFIESTA (new Forest Inventory ESTimation and Analysis) Estimation methods used within the Diabolo T2.3.1 case study.* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (the EU's Horizon 2020 programme).

Bouriaud, O., Brion, P., Chauvet, G., Duong, T.H.K., & Pulkkinen, M. 2024. The weight share method in forest inventories: refining the relation between points and trees. *Canadian journal of forest research*, **54**(10), 1129–1141.

Breidt, F. Jay, & Opsomer, Jean D. 2008. Endogenous post-stratification in surveys: Classifying with a sample-fitted model. *The annals of statistics*, **36**(1), 403–427.

Casella, George, & Berger, Roger L. 2002. *Statistical inference.* 2nd edn. Pacific Grove, CA: Duxbury.

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

Dahlke, Mark, Breidt, F. Jay, Opsomer, Jean D., & Van Keilegom, Ingrid. 2013. Nonparametric endogenous post-stratification estimation. *Statistica sinica*, **23**(1), 189–211.

Ene, L., T., Adrian, L., Adolt, R., & Fejfar, J. 2018. *Deliverable d2.10 - partial report on the development of imputation techniques and updating algorithms, including concept and implementation.* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (DIABOLO).

Estevao, Victor M., & Särndal, Carl-Erik. 2004. The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of official statistics*, **20**(1), 129–145.

INSPIRE Maintenance and Implementation Group (MIG). 2023 (Jan.). *INSPIRE data specification on geographical grid systems – technical guidelines (d2.8.i.2).* Tech. rept. European Commission. Version 3.2.0 (released 31 January 2023).

Mandallaz, D. 1991. *A unified approach to sampling theory for forest inventory based on infinite population and superpopulation models.* Ph.D. thesis, Swiss Federal Institute of Technology (ETH), Zurich.

Massey, Alexander, & Lanz, Adrian. 2014. Integrating remote sensing and past inventory data under the new annual design of the swiss national forest inventory using three-phase design-based regression estimation. *Canadian journal of forest research*, **44**(06), 1177–1186.

McRoberts, Ronald E., Næsset, Erik, Gobakken, Terje, & Bollandsås, Ole Martin. 2015. Indirect and direct estimation of forest biomass change using forest inventory and airborne laser scanning data. *Remote sensing of environment*, **164**, 36–42.

Miettinen, Jukka, Adolt, Radim, Breidenbach, Johannes, Fejfar, Jiří, Hanáková, Jana, Kohn, Ivo, Kratěna, Lukas, Myllymäki, Mari, Seitsonen, Lauri, Tergujeff, Renne, & Závodský, Jiří. 2024. *Pathfinder d2.1: Initial characteristics for mapping and estimation platform.* Deliverable D2.1. PathFinder Project, Horizon 2020, European Commission, Brussels, Belgium. Contributions: VTT (Forestry TEP sections); ÚHÚL (nFIESTA sections); NIBIO (overview, advice, supervision); LUKE (overview, advice, supervision).

Schadauer, Klemens, Astrup, Rasmus, Breidenbach, Johannes, Fridman, Jonas, Gräber, Stephan, Köhl, Michael, Korhonen, Kari T., Johannsen, Vivian Kvist, Morneau, Francois, Päivinen, Risto, & Riedel, Thomas. 2024. Access to exact national forest inventory plot locations must be carefully evaluated. *New phytologist*, **242**(2), 347–350.

Särndal, C. E., Swensson, B., & Wretman, J. 2003. *Model assisted survey sampling.* Springer.

Tipton, Jennifer, Opsomer, Jean, & Moisen, Gretchen. 2013. Properties of endogenous post-stratified estimation using remote sensing data. *Remote sensing of environment*, **139**, 130–137.

Annexes

A Derivation of variances of regression estimators of total

A.1 Variance of single-phase regression estimator of the total with exhaustive auxiliaries

The single-phase regression estimator is defined

$$\hat{t}_{y,1p} = \mathbf{I_D \Pi_D Y'_D} + \Delta'_{z^{(0)}} \mathbf{G}_{\beta_+^{(0)}} \mathbf{\Pi}_+ \mathbf{Y}'_+ \tag{67}$$

The symbols in this equation are explained in Sections 2.1 and 2.2.

This estimator also equals the total of predictions on the bases of estimated model coefficients $\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})}$ and the estimated total of (empirical) residuals in D:

$$\hat{t}_{y,1pm} = \int_{\mathcal{U}} \mathbf{Z_D^{(0)'}}(x) \hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} dx + \sum_{x \in s_2} \frac{e^{(D)}(x)}{\pi(x)} \tag{68}$$

$$= \mathbf{t_{z^{(0)}}} \hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} + \hat{t}_e \tag{69}$$

The total estimator of empirical residuals $\hat{t}_e$ can be expressed as differences between the observed values of the target variable $y(x)$ and the prediction of the working model

$$\hat{t}_e = \sum_{x \in s_2} \frac{y^{(D)}(x) - \mathbf{Z_D^{(0)'}}(x) \hat{\boldsymbol{\beta}}_+^{(\mathbf{0})}}{\pi(x)}. \tag{70}$$

Theoretical (linear) model of relation between the target variable $y(x)$ and the auxiliary vector $\mathbf{Z_D^{(0)'}}(x)$ is defined

$$y^{(D)}(x) = \mathbf{Z_D^{(0)'}}(x) \boldsymbol{\beta}^{(0)} + E(x), \tag{71}$$

Where $\mathbf{Z_D^{(0)'}}(x)\boldsymbol{\beta}^{(0)}$ is the theoretical prediction, $\mathbf{Z_D^{(0)}}(x)$ is a vector of the p auxiliaries with dimension $p \times 1$, and $E(x)$ is the theoretical residual of the true model with parameter vector $\boldsymbol{\beta}^{(0)}$ of dimension $p \times 1$ defined by

$$\boldsymbol{\beta}^{(0)} = \mathbf{A^{-1}U}. \tag{72}$$

The matrix A of dimension $p_0 \times p_0$ is defined as a definite integral in the entire region D_+, where the model is parameterised:

$$\mathbf{A} = \int_{D_+} \mathbf{Z_+^{(0)}}(x) \mathbf{\Sigma}(x) \mathbf{Z_+^{(0)'}}(x) dx. \tag{73}$$

The matrix $\mathbf{U}$ of dimension $p \times 1$ is also defined as a definite integral in D_+:

$$\mathbf{U} = \int_{D_+} \mathbf{Z_+^{(0)}}(x) \mathbf{\Sigma}(x) y(x) dx, \tag{74}$$

where $y(x)$ is the value of the target variable observed at the point x in the parametrization region D_+.

In practical applications, the coefficients of the theoretical model are not available and must be estimated. So, we replace the theoretical model by its working version:

$$y^D(x) = \mathbf{Z_D^{(0)}}'(x)\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} + e(x), \tag{75}$$

where $\hat{\boldsymbol{\beta}}_+$ is the vector of model coefficients estimated from sample data, and $e(x)$ is the empirical residual. Model parameters are estimated by

$$\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} = \hat{\mathbf{A}}_+^{-\mathbf{1}}\mathbf{U}_+, \tag{76}$$

where

$$\hat{\mathbf{U}}_+ = \mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{Y}'_+ \qquad \hat{\mathbf{A}}_+ = \mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{Z}_+^{(\mathbf{0})}, \tag{77}$$

are estimators of $\mathbf{U}$ and $\mathbf{A}$ matrices. The vector $\mathbf{Y}_+$ contains values of the target variable $y^{(D_+)}(x)$ observed on sample points, its dimension is $1 \times n_{2,+}$ ($n_{2,+}$ corresponding to the number of sample points in parameterisation region D_+) and the matrix $\mathbf{Z}_+^{(\mathbf{0})}$ of auxiliaries defined in section 2.2 is of dimension $p_0 \times n_{2,+}$.

The actual variance of regression estimator includes the variance of the model parameters $\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})}$, that is, the variance of the differences between the true and estimated model parameters $\left[\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \boldsymbol{\beta}^{(0)}\right]$. These differences can be approximated by first-order derivatives of $\boldsymbol{\beta}$ (Taylor approximation). Knowing that

$$\mathbf{A^{-1}A} = \mathbf{I}, \tag{78}$$

and

$$d(\mathbf{A^{-1}A}) = d(\mathbf{I}) = 0 \tag{79}$$

$$d(\mathbf{A^{-1}})\mathbf{A} + \mathbf{A^{-1}}d(\mathbf{A}) = 0 \tag{80}$$

$$d(\mathbf{A^{-1}})\mathbf{A} = -\mathbf{A^{-1}}d(\mathbf{A}) \tag{81}$$

$$d(\mathbf{A^{-1}}) = -\mathbf{A^{-1}}d(\mathbf{A})\mathbf{A}^{-1} \tag{82}$$

the first derivation of the parameter $\boldsymbol{\beta}$ according to (76) is

$$d\left[\boldsymbol{\beta}^{(0)}\right] = d(\mathbf{A}^{-1})\mathbf{U} + \mathbf{A}^{-1}d(\mathbf{U}) \tag{83}$$

The difference between the true and estimated vector of model parameters can be approximated by

$$\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \boldsymbol{\beta}^{(0)} \approx -\mathbf{A}^{-1}\left[\hat{\mathbf{A}}_+ - \mathbf{A}\right]\mathbf{A}^{-1}\mathbf{U} + \mathbf{A}^{-1}\left[\hat{\mathbf{U}}_+ - \mathbf{U}\right] \tag{84}$$

$$\approx \left[\mathbf{A}^{-1}\mathbf{A} - \mathbf{A}^{-1}\hat{\mathbf{A}}_+\right]\boldsymbol{\beta}^{(0)} + \mathbf{A}^{-1}\hat{\mathbf{U}}_+ - \boldsymbol{\beta}^{(0)} \tag{85}$$

$$\approx \mathbf{A}^{-1}\left[\hat{\mathbf{U}}_+ - \hat{\mathbf{A}}_+\right] \tag{86}$$

$$\approx \mathbf{A}^{-1}\left[\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{Y}'_+ - \mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\beta}^{(0)}\right] \tag{87}$$

$$\approx \mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{E}_+ \tag{88}$$

where $\mathbf{E}'_+$ is $n_+ \times 1$ vector of theoretical residuals $E^{(D_+)}(x)$ of the true model.

Reformulating the total estimator of empirical residuals in terms of the parameter differences, we get

$$\hat{t}_e = \sum_{x \in s_2} \frac{y(x) - \mathbf{Z_D^{(0)}}'(x)\boldsymbol{\beta} - \mathbf{Z_D^{(0)}}'(x)\left[\hat{\boldsymbol{\beta}}_+^{\mathbf{0}} - \boldsymbol{\beta}^{(0)}\right]}{\pi(x)+} \tag{89}$$

$$\approx \sum_{x \in s_2} \frac{E(x) - \mathbf{Z_D^{(0)}}'(x)\mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Pi}_+\mathbf{E}'_+}{\pi(x)+} \tag{90}$$

$$\approx \hat{t}_E - \hat{\mathbf{t}}_{\mathbf{s_2 z^{(0)}}}\mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Pi}_+\mathbf{E}'_+ \tag{91}$$

Similarly, for the true total of empirical predictions we can write

$$\mathbf{t}_{\mathbf{z}^{(0)}}\hat{\boldsymbol{\beta}}_+ = \int_{\mathcal{U}} \mathbf{Z_D^{(0)}}'(x)\boldsymbol{\beta}^{(0)}dx + \int_{\mathcal{U}} \mathbf{Z_D^{(0)}}'(x)\left[\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \boldsymbol{\beta}^{(0)}\right] \tag{92}$$

$$\approx t_{\tilde{y}_D} + \mathbf{t}_{\mathbf{z}^{(0)}}\mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Pi}_+\mathbf{E}'_+. \tag{93}$$

By combining (91) and (93) we get the following approximation of single-phase regression estimator

$$\hat{t}_{y,1pm} \approx t_{\tilde{y}_D} + \mathbf{t}_{\mathbf{z}^{(0)}}\mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{E}'_+ + \hat{t}_E - \hat{\mathbf{t}}_{\mathbf{s_2 z^{(0)}}}\mathbf{A}^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{E}'_+ \tag{94}$$

$$\approx t_{\tilde{y}_D} + \hat{t}_E + (\mathbf{t}_{\mathbf{z}^{(0)}} - \hat{\mathbf{t}}_{\mathbf{s_2 z^{(0)}}})'\hat{\mathbf{A}}_+^{-1}\mathbf{Z}_+^{(\mathbf{0})}\boldsymbol{\Sigma}_+\boldsymbol{\Pi}_+\mathbf{E}'_+ \tag{95}$$

$$\approx t_{\tilde{y}_D} + \hat{t}_E + \boldsymbol{\Delta}_{\mathbf{z}^{(\mathbf{0})}}{}'\tilde{\mathbf{G}}_{\beta_+^{(0)}}\boldsymbol{\Pi}_+\mathbf{E}'_+ \tag{96}$$

$$\approx t_{\tilde{y}_D} + \sum_{x \in s} \frac{\phi(x)}{\pi(x)}, \tag{97}$$

$$\approx t_{\tilde{y}_D} + \hat{t}_\phi \tag{98}$$

where

$$\phi(x) = e^{(D)}(x) + \boldsymbol{\Delta}_{\mathbf{z}^{(\mathbf{0})}}\tilde{\mathbf{G}}_{\beta_+^{(0)}}(x)e^{(D_+)}(x) \tag{99}$$

and its terms defined in Section 2.2.

The variance of $\hat{t}_{y,1pm}$ is the variance of $\hat{t}_\phi$ and can be estimated by HTC, see formula (31). The first term of (98), the true sum of empirical predictions, is a constant, so it does not contribute to variance.

If no clusters but single plots are used ($m = 1$), then $e^{(D)}(x) = e^{(D_+)}(x) = e(x)$ and the single-phase regression estimator is defined

$$\hat{t}_{y,1pm} \approx t_{\tilde{y}_D} + \sum_{x \in s2} \frac{\left[I_D(x) + \hat{\mathbf{A}}_+^{-1}\mathbf{Z}_+^{(\mathbf{0})}(x)\sigma^{-2}(x)\right]e(x)}{\pi(x)}, \tag{100}$$

$$\approx t_{\tilde{y}_D} + \sum_{x \in s2} \frac{g(x)e(x)}{\pi(x)}, \tag{101}$$

with its variance matching the HTC variance for the total of the product of g-weights and residuals. Note that in the definitions of $\tilde{\mathbf{G}}_{\beta_+^{(0)}}(x)$ and $g(x)$ we use the estimated matrix $\hat{\mathbf{A}}_+^{-1}$ instead of the true matrix $\mathbf{A}^{-1}$. The estimation of the variance also uses empirical residuals instead of the unknown errors of the true model.

A.2 Variance of two-phase regression estimator of the total without exhaustive auxiliaries

This estimator was introduced in Section 2.1. It can also be defined as the sum of predictions plus a correction by the sum of residuals. However, the sum of predictions is not known but is estimated using the first-phase sample

$$\hat{t}_{y,2p} = \hat{\mathbf{t}}_{\mathbf{z}^{(1)}} \hat{\boldsymbol{\beta}}_{+}^{(\mathbf{1})} + \hat{t}_{\dot{e}} \tag{102}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \hat{\boldsymbol{\beta}}_{+}^{(\mathbf{1})}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{e}^{(D)}(x)}{\pi^*(x)} \tag{103}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \hat{\boldsymbol{\beta}}_{+}^{(\mathbf{1})}}{\pi_1(x)} + \sum_{x \in s_2} \frac{y^{(D)}(x) - \mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \hat{\boldsymbol{\beta}}_{+}^{(\mathbf{1})}}{\pi^*(x)}. \tag{104}$$

Note that for simplicity and convenience in typing, the model parameters and their estimates use the same symbols as in the previous annex. However, they are defined using the auxiliary vectors $\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})}$ instead of $\mathbf{Z}_{\mathbf{D}}^{(\mathbf{0})}$. The estimator can be transformed into a form using the differences between the estimated and true model parameters:

$$\hat{t}_{y,2p} = \underbrace{\sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \boldsymbol{\beta}^{(1)}}{\pi_1(x)}}_{\text{prediction of true model}} + \underbrace{\sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \left[\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)}\right]}{\pi_1(x)}}_{\substack{\text{error due to estimated} \\ \text{model parameters}}} +$$

$$+ \sum_{x \in s_2} \frac{y^{(D)}(x) - \mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \hat{\boldsymbol{\beta}}_{+}^{(1)}}{\pi^*(x)} \tag{105}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{E}^{(D)}(x)}{\pi^*(x)}$$

$$+ \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \left[\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)}\right]}{\pi_1(x)} - \sum_{x \in s_2} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \left[\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)}\right]}{\pi^*(x)} \tag{106}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{E}^{(D)}(x)}{\pi_2(x)} + \left[\hat{\mathbf{t}}_{\mathbf{s_1}, \mathbf{z}_{\mathbf{D}}^{(\mathbf{1})}} - \hat{\mathbf{t}}_{\mathbf{s_2}, \mathbf{z}_{\mathbf{D}}^{(\mathbf{1})}}\right]' \left[\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)}\right] \tag{107}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z}_{\mathbf{D}}^{(\mathbf{1})'}(x) \boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{E}^{(D)}(x)}{\pi_2(x)} + \boldsymbol{\Delta}_{\mathbf{z}^{(1)}}{}' \left[\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)}\right] \tag{108}$$

in analogy to the approximation derived in Annex A.1

$$\hat{\boldsymbol{\beta}}_{+}^{(1)} - \boldsymbol{\beta}^{(1)} \approx \mathbb{A}^{-1} \mathbf{Z}_{+}^{(\mathbf{1})} \boldsymbol{\Sigma}_{+} \boldsymbol{\Pi}_{+} \dot{\mathbf{E}}_{+}, \tag{109}$$

where the matrix $\mathbb{A}$ and its estimator are now redefined by using $\mathbf{Z}_{+}^{(\mathbf{1})}$ auxiliaries

$$\mathbb{A} = \int_{D_+} \mathbf{Z}_{+}^{(\mathbf{1})}(x) \boldsymbol{\Sigma}(x) \mathbf{Z}_{+}^{(\mathbf{1})'}(x) dx, \qquad \hat{\mathbb{A}}_{+} = \mathbf{Z}_{+}^{(\mathbf{1})} \boldsymbol{\Sigma}_{+} \boldsymbol{\Pi}_{+} \mathbf{Z}_{+}^{(\mathbf{1})}. \tag{110}$$

The meaning of symbols is defined in Section 2.1.

The two-phase regression estimator is rewritten as

$$\hat{t}_{y,2p} \approx \sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{E}^{(D)}(x)}{\pi^*(x)} + \boldsymbol{\Delta}_{\mathbf{z^{(1)}}}{}' \mathbb{A}^{-1} \mathbf{Z}_+^{(\mathbf{1})} \boldsymbol{\Sigma}_+ \boldsymbol{\Pi}_+ \dot{\mathbf{E}}_+, \tag{111}$$

$$\approx \sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\dot{E}^{(D)}(x)}{\pi^*(x)} + \boldsymbol{\Delta}_{\mathbf{z^{(1)}}}{}' \hat{\mathbb{A}}_+^{-1} \mathbf{Z}_+^{(\mathbf{1})} \boldsymbol{\Sigma}_+ \boldsymbol{\Pi}_+ \dot{\mathbf{E}}_+ \tag{112}$$

$$\approx \sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_2} \frac{\rho(x)}{\pi^*(x)}, \tag{113}$$

where

$$\rho(x) = \dot{e}^{(D)}(x) + \boldsymbol{\Delta}_{\mathbf{z^{(1)}}} \tilde{\mathbf{G}}_{\beta_+^{(1)}}(x) \dot{e}^{(D_+)}(x) \tag{114}$$

and its terms defined in Section 2.1.

The variance of any two-phase estimator $\hat{\theta}_{2p}$ is defined

$$\mathbb{V}(\hat{\theta}_{2p}) = \mathbb{V}_1 \left[\mathbb{E}_{2|1}(\hat{\theta}_{2p}) \right] + \mathbb{E}_1 \left[\mathbb{V}_{2|1}(\hat{\theta}_{2p}) \right], \tag{115}$$

The conditional expectation of $\hat{t}_{y,2p}$ based on (112)

$$\mathbb{E}_{2|1}\left[\hat{t}_{y,2p}\right] = \mathbb{E}_{2|1}\left[\sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)}\right] + \mathbb{E}_{2|1}\left[\sum_{x \in s_2} \frac{y^{(D)}(x) - \mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi^*(x)}\right]$$
$$+ \mathbb{E}_{2|1}\left[\boldsymbol{\Delta}_{\mathbf{z^{(1)}}}{}' \mathbb{A}^{-1} \mathbf{Z}_+^{(\mathbf{0})} \boldsymbol{\Sigma}_+ \boldsymbol{\Pi}_+ \dot{\mathbf{E}}_+\right] \tag{116}$$

$$= \sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)} - \sum_{x \in s_1} \frac{\mathbf{Z_D^{(1)}}'(x)\boldsymbol{\beta}^{(1)}}{\pi_1(x)} + \sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)} + 0 \tag{117}$$

$$= \sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)}. \tag{118}$$

The conditional expectation of the third term in (112) is zero because the conditional expectation of $\boldsymbol{\Delta}_{\mathbf{z^{(1)}}} = \hat{\mathbf{t}}_{\mathbf{s_1}, \mathbf{z_D^{(1)}}} - \hat{\mathbf{t}}_{\mathbf{s_2}, \mathbf{z_D^{(1)}}}$ is zero.

The conditional variance of (113) is simply the following

$$\mathbb{V}_{2|1}\left[\hat{t}_{y,2p}\right] = \mathbb{V}_{2|1}\left[\sum_{x \in s_2} \frac{\rho(x)}{\pi^*(x)}\right], \tag{119}$$

because the first term does not vary if the first phase sample is kept constant.

The variance of the conditional expectation is equal to

$$\mathbb{V}_1\left\{\mathbb{E}_{2|1}\left[\hat{t}_{y,2p}\right]\right\} = \mathbb{V}_1\left[\sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)}\right] \tag{120}$$

$$= \mathbb{E}_1\left[\sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)} \sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)}\right]$$
$$- \mathbb{E}_1\left[\sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)}\right] \mathbb{E}_1\left[\sum_{x \in s_1} \frac{y^{(D)}(x)}{\pi_1(x)}\right] \tag{121}$$

where for the first term, we have

$$\mathbb{E}_1\left[\sum_{x\in s_1}\frac{y^{(D)}(x)}{\pi_1(x)}\sum_{x\in s_1}\frac{y^{(D)}(x)}{\pi_1(x)}\right]=\mathbb{E}_1\left[\sum_{i=1}^{n_1}\frac{[y^{(D)}(x_i)]^2}{\pi_1^2(x_i)}\right] \tag{122}$$

$$+\mathbb{E}_1\left[\sum_{i=1}^{n_1}\sum_{\substack{j=1\\j\neq i}}^{n_1}\frac{y^{(D)}(x_i)y^{(D)}(x_i)}{\pi_1(x_i)\pi_1(x_j)}\right] \tag{123}$$

The first term of (123) equals

$$\begin{aligned}&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1}\left[\frac{y^{(D)}(x_i)}{\pi_1(x_i)}\right]^2\right\}f_1(x_1,x_2,\dots x_{n_1})dx_1dx_2\dots dx_{n_1}\\&=\sum_{i=1}^{n_1}\int_{\mathcal{U}}\left[\frac{y^{(D)}(x)}{\pi_1(x)}\right]^2 f_{1,i}(x)dx=\int_{\mathcal{U}}\left[\frac{y^{(D)}(x)}{\pi_1(x)}\right]^2\sum_{i=1}^{n_1}f_{1,i}(x)dx\\&=\int_{\mathcal{U}}\left[\frac{y^{(D)}(x)}{\pi_1(x)}\right]^2\pi_1(x)dx=\int_{\mathcal{U}}\frac{[y^{(D)}(x)]^2}{\pi_1(x)}dx\end{aligned} \tag{124}$$

where the joint probability density $f_1(x_1,x_2,\dots x_{n_1})$ refers to simultaneous selection of all sample points of s_1 in the first phase, and the probability density $f_{1,i}(x)$ refers to the selection of point x in the i-th draw. The transition from multiple to single integral is possible according to Fubini's theorem, which allows for iterative integration over each x_i in every step, so the number of points in the joint probability density is reduced by one in each iteration until only one point remains. The final operation uses the definition of inclusion density in

$$\pi_1(x)=\sum_{i=1}^{n_1}f_{1,i}(x). \tag{125}$$

The second term of (123) is defined

$$\begin{aligned}&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1}\sum_{\substack{x_j\in s_1\\x_j\neq x_i}}\frac{y^{(D)}(x_i)y^{(D)}(x_j)}{\pi_1(x_i)\pi_1(x_j)}\right\}f_1(x_1,x_2,\dots x_{n_1})dx_1dx_2\dots dx_{n_1}\\&=\sum_{i=1}^{n_1}\sum_{\substack{j=1\\j\neq i}}^{n_1}\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y^{(D)}(x)y^{(D)}(x')}{\pi_1(x)\pi_1(x')}f_{1,i,j}(x,x')dxdx'\\&=\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y^{(D)}(x)y^{(D)}(x')}{\pi_1(x)\pi_1(x')}\sum_{i=1}^{n_1}\sum_{\substack{j=1\\j\neq i}}^{n_1}f_{1,i,j}(x,x')dxdx'\\&=\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y^{(D)}(x)y^{(D)}(x')}{\pi_1(x)\pi_1(x')}\pi_1(x,x')dxdx'.\end{aligned} \tag{126}$$

Note that the double integral, somewhat surprisingly, includes the region where the sample points x and x' are identical. This region is a line in the 2D space $\mathcal{U}\times\mathcal{U}$, so its measure is zero. Thus, it has no effect on the value of the double integral, which is actually identical to the second expectation term of (123).

The last term in (121) equals

$$\begin{aligned}
&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\sum_{x_i\in s_1}\left\{\frac{y^{(D)}(x_i)}{\pi_1(x_i)}\right\}f_1(x_1,x_2,\dots x_{n_1})dx_1dx_2\dots dx_{n_1}\times\\
&\quad\times\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\sum_{x_i\in s_1}\left\{\frac{y^{(D)}(x_j)}{\pi_1(x_j)}\right\}f_1(x_1,x_2,\dots x_{n_1})dx_1dx_2\dots dx_{n_1}\\
&\quad=\int_{\mathcal{U}}\frac{y^{(D)}(x)}{\pi_1(x)}\sum_{i=1}^{n_1}f_{1,i}(x)dx\int_{\mathcal{U}}\frac{y^{(D)}(x)}{\pi_1(x)}\sum_{i=1}^{n_1}f_{1,i}(x)dx \qquad (127)\\
&\quad=\int_{\mathcal{U}}\frac{y^{(D)}(x)}{\pi_1(x)}\pi_1(x)dx\int_{\mathcal{U}}\frac{y^{(D)}(x)}{\pi_1(x)}\pi_1(x)dx\\
&\quad=\int_D y^{(D)}(x)dx\int_{\mathcal{U}}y^{(D)}(x)dx=\int\int_{\mathcal{U}\times\mathcal{U}}y^{(D)}(x)y^{(D)}(x')dxdx'
\end{aligned}$$

The variance of the conditional expectations is composed by the three parts that have been derived above

$$\begin{aligned}
\mathbb{V}_1\left\{\mathbb{E}_{2|1}\left[\hat{t}_{y,2p}\right]\right\}&=\int_{\mathcal{U}}\frac{[y^{(D)}(x)]^2}{\pi_1(x)}d_x+\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y^{(D)}(x)y^{(D)}(x')}{\pi_1(x)\pi_1(x')}\pi_1(x,x')d_xd_{x'}\\
&\quad-\int\int_{\mathcal{U}\times\mathcal{U}}y^{(D)}(x)y^{(D)}(x')dxdx' \qquad (128)\\
&=\int_{\mathcal{U}}\frac{[y^{(D)}(x)]^2}{\pi_1(x)}d_x\\
&\quad+\int\int_{\mathcal{U}\times\mathcal{U}}y^{(D)}(x)y^{(D)}(x')\frac{\pi_1(x,x')-\pi_1(x)\pi_1(x')}{\pi_1(x)\pi_1(x')}d_xd_{x'} \qquad (129)
\end{aligned}$$

Its estimator is obtained from the second phase sample as the HTC total estimator. It means that integrals are replaced by sums and double sums, and the integrands are divided by inclusion density or pairwise density including both sampling phases.

$$\hat{\mathbb{V}}_1\left\{\mathbb{E}_1\left[\hat{t}_{y,2p}\right]\right\}=\sum_{x\in s_2}\frac{[y^{(D)}(x)]^2}{\pi^*(x)\pi_1(x)}+\sum_{x\in s_2}\sum_{\substack{x'\in s_2\\x'\neq x}}\frac{\pi_1(x,x')-\pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} \qquad (130)$$

Conditional variance is defined with respect to sampling from the finite population of the first-phase points

$$\begin{aligned}
\mathbb{V}_{2|1}\left[\hat{t}_{y,2p}\right]&=\sum_{x_i\in s_1}\sum_{x_j\in s_1}\rho(x_i)\rho(x_j)\frac{\pi_{2|1}(x_i,x_j)-\pi_{2|1}(x_i)\pi_{2|1}(x_j)}{\pi^*(x_i)\pi^*(x_j)} \qquad (131)\\
&=\sum_{x\in s_1}\frac{\rho^2(x)}{\pi_1(x)\pi^*(x)}[1-\pi_{2|1}(x)]\\
&\quad+\sum_{x_i\in s_1}\sum_{\substack{x_j\in s_1\\x_j\neq x_i}}\rho(x_i)\rho(x_j)\frac{\pi_{2|1}(x_i,x_j)-\pi_{2|1}(x_i)\pi_{2|1}(x_j)}{\pi^*(x_i)\pi^*(x_j)} \qquad (132)
\end{aligned}$$

Note that the first term of (132) was obtained knowing that $\pi_{2|1}(x_i,x_j)=\pi_{2|1}(x)$ if $x_i=x_j=x$.

The expectation of the conditional variance over the set of all first-phase samples equals

$$\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1}\sum_{x_j\in s_1}\rho(x_i)\rho(x_j)\frac{\pi_{2|1}(x_i,x_j)-\pi_{2|1}(x_i)\pi_{2|1}(x_j)}{\pi^*(x_i)\pi^*(x_j)}\right\}\times \\ \times f_1(x_1,x_2\ldots x_{n_1})dx_1dx_2\ldots dx_{n_1} \tag{133}$$

and is simplified to

$$\int\int_{\mathcal{U}\times\mathcal{U}}\rho(x)\rho(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x')}{\pi^*(x)\pi^*(x')}\pi_1(x,x')dxdx' \tag{134}$$

It is estimated based on second-phase sample as follows

$$\hat{\mathbb{E}}_1\left\{\mathbb{V}_{2|1}\left[\hat{t}_{y,2p}\right]\right\}=\sum_{x_i\in s_2}\sum_{x_j\in s_2}\rho(x)\rho(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x')}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x')} \tag{135}$$

$$=\sum_{x\in s_2}\left[\frac{\rho(x)}{\pi^*(x)}\right]^2[1-\pi_{2|1}(x)] \tag{136}$$

$$+\sum_{x\in s_2}\sum_{\substack{x'\in s_2\\x'\neq x}}\rho(x)\rho(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x)}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x)} \tag{137}$$

Finally, the variance of $\hat{t}_{y,2p}$ is estimated by adding (137) to (130)

$$\hat{\mathbb{V}}[\hat{t}_{y,2p}]=\sum_{x\in s_2}\frac{[y^{(D)}(x)]^2}{\pi^*(x)\pi_1(x)}+\sum_{x\in s_2}\sum_{\substack{x'\in s_2\\x'\neq x}}\frac{\pi_1(x,x')-\pi_1(x)\pi_1(x')}{\pi^*(x,x')\pi_1(x)\pi_1(x')} \\ +\sum_{x\in s_2}\left[\frac{\rho(x)}{\pi^*(x)}\right]^2[1-\pi_{2|1}(x)] \tag{138}$$

$$+\sum_{x\in s_2}\sum_{\substack{x'\in s_2\\x'\neq x}}\rho(x)\rho(x')\frac{\pi_{2|1}(x,x')-\pi_{2|1}(x)\pi_{2|1}(x)}{\pi_{2|1}(x,x')\pi^*(x)\pi^*(x)}$$

A.3 Variance of two-phase regression estimator of the total with exhaustive auxiliaries

This estimator is defined in Section 2.2. It can be reformulated as follows,

$$\begin{aligned}
\hat{t}_{y,2pm} &= \sum_{x\in s_2} \frac{y^{(D)}(x)}{\pi^{(*)}(x)} + \sum_{x\in s_1} \frac{\mathbf{Z_D}'(x)\hat{\boldsymbol{\beta}}_+}{\pi_1(x)} - \sum_{x\in s_2} \frac{\mathbf{Z_D}'(x)\hat{\boldsymbol{\beta}}_+}{\pi^*(x)} \\
&\quad + \int_{\mathcal{U}} \mathbf{Z_D^{(0)}}'(x)\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} dx - \sum_{x\in s_1} \frac{\mathbf{Z_D^{(0)}}'(x)\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})}}{\pi_1(x)} \qquad (139) \\
&= \sum_{x\in s_2} \frac{y^{(D)}(x)}{\pi^*(x)} + \sum_{x\in s_1} \frac{y^{(D)}(x) - \ddot{e}^{(D)}(x)}{\pi_1(x)} \\
&\quad - \sum_{x\in s_2} \frac{y^{(D)}(x) - \ddot{e}^{(D)}(x)}{\pi^*(x)} + \mathbf{t_{z^{(0)}}}\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \sum_{x\in s_1} \frac{y^{(D)}(x) - e^{(D)}(x)}{\pi_1(x)} \qquad (140) \\
&= \mathbf{t_{z^{(0)}}}\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \sum_{x\in s_1} \frac{\ddot{e}^{(D)}(x)}{\pi_1(x)} + \sum_{x\in s_2} \frac{\ddot{e}^{(D)}(x)}{\pi^*(x)} + \sum_{x\in s_1} \frac{e^{(D)}(x)}{\pi_1(x)} \qquad (141)
\end{aligned}$$

The residuals can be expressed as differences between observed $y^{(D)}(x)$ and the corresponding predictions

$$\begin{aligned}
\hat{t}_{y,2pm} &= \mathbf{t_{z^{(0)}}}\hat{\boldsymbol{\beta}}_+ - \sum_{x\in s_1} \frac{y^{(D)}(x) - \mathbf{Z_D}'(x)\hat{\boldsymbol{\beta}}_+}{\pi_1(x)} \qquad (142) \\
&\quad + \sum_{x\in s_2} \frac{y^{(D)}(x) - \mathbf{Z_D}'(x)\hat{\boldsymbol{\beta}}_+}{\pi^*(x)} + \sum_{x\in s_1} \frac{y^{(D)}(x) - \mathbf{Z_D^{(0)}}'(x)\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})}}{\pi_1(x)} \qquad (143)
\end{aligned}$$

The next step introduces differences between the estimated and true model coefficients as well as the theoretical residuals

$$\begin{aligned}
\hat{t}_{y,2pm} &= \mathbf{t_{z^{(0)}}}\boldsymbol{\beta} + \mathbf{t_{z^{(0)}}}\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right] - \sum_{x\in s_1} \frac{y^{(D)}(x) - \mathbf{Z_D}'(x)\boldsymbol{\beta} - \mathbf{Z_D}'(x)\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right]}{\pi_1(x)} \\
&\quad + \sum_{x\in s_2} \frac{y^{(D)}(x) - \mathbf{Z_D}'(x)\boldsymbol{\beta} - \mathbf{Z_D}'(x)\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right]}{\pi^*(x)} \\
&\quad + \sum_{x\in s_1} \frac{y^{(D)}(x) - \mathbf{Z_D^{(0)}}'(x)\boldsymbol{\beta} - \mathbf{Z_D^{(0)}}'(x)\left[\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \boldsymbol{\beta}^{(0)}\right]}{\pi_1(x)} \qquad (144) \\
&= \mathbf{t_{z^{(0)}}}\boldsymbol{\beta} - \mathbf{t_{z^{(0)}}}\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right] - \sum_{x\in s_1} \frac{\ddot{E}^{(D)}(x) - \mathbf{Z_D}'(x)\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right]}{\pi_1(x)} \\
&\quad + \sum_{x\in s_2} \frac{\ddot{E}^{(D)}(x) - \mathbf{Z_D}'(x)\left[\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}\right]}{\pi^*(x)} \\
&\quad + \sum_{x\in s_1} \frac{E^{(D)}(x) - \mathbf{Z_D^{(0)}}'(x)\left[\hat{\boldsymbol{\beta}}_+^{(\mathbf{0})} - \boldsymbol{\beta}^{(0)}\right]}{\pi_1(x)} \qquad (145)
\end{aligned}$$

Next the approximation derived in Annex A.1

$$\hat{\boldsymbol{\beta}}_{+}^{(0)} - \boldsymbol{\beta}^{(0)} \approx \mathbf{A}^{-1}\mathbf{Z}_{+}^{(\mathbf{0})}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{E}_{+}, \tag{146}$$

and an analogous approximation

$$\hat{\boldsymbol{\beta}}_{+} - \boldsymbol{\beta} \approx \mathcal{A}^{-1}\mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\ddot{\mathbf{E}}_{+}, \tag{147}$$

enter the derivations.

The matrix $\mathcal{A}$ and its estimator are now defined by using $\mathcal{Z}_{+}$ auxiliaries

$$\mathcal{A} = \int_{D_{+}} \mathbf{Z}_{+}(x)\boldsymbol{\Sigma}_{+}(x)\mathbf{Z}_{+}{}'(x)dx. \qquad \hat{\mathcal{A}}_{+} = \mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{Z}_{+}, \tag{148}$$

The meaning of symbols is defined in Section 2.2.

Then, the two-phase regression estimator with exhaustive auxiliaries is approximated by

$$\begin{aligned}\hat{t}_{y,2pm} \approx\ & \mathbf{t}_{\mathbf{z}^{(0)}}\boldsymbol{\beta} + \mathbf{t}_{\mathbf{z}^{(0)}}\mathbf{A}^{-1}\mathbf{Z}_{+}^{(\mathbf{0})}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{E}_{+} \\ & - \sum_{x\in s_1} \frac{\ddot{E}^{(D)}(x) - \mathbf{Z}_{\mathbf{D}}{}'(x)\mathcal{A}^{-1}\mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\ddot{\mathbf{E}}_{+}}{\pi_1(x)} \\ & + \sum_{x\in s_2} \frac{\ddot{E}^{(D)}(x) - \mathbf{Z}_{\mathbf{D}}{}'(x)\mathcal{A}^{-1}\mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\ddot{\mathbf{E}}_{+}}{\pi^{*}(x)} \\ & + \sum_{x\in s_1} \frac{E^{(D)}(x) - \mathbf{Z}_{\mathbf{D}}^{(\mathbf{0})'}(x)\mathbf{A}^{-1}\mathbf{Z}_{+}^{(\mathbf{0})}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{E}_{+}}{\pi_1(x)},\end{aligned} \tag{149}$$

which can be rearranged as follows

$$\begin{aligned}\hat{t}_{y,2pm} \approx\ & \mathbf{t}_{\mathbf{z}^{(0)}}\boldsymbol{\beta} + \sum_{x\in s_1}\frac{E^{(D)}(x)}{\pi_1(x)} + \sum_{x\in s_2}\frac{\ddot{E}^{(D)}(x)}{\pi^{*}(x)} - \sum_{x\in s_1}\frac{\ddot{E}^{(D)}(x)}{\pi_1(x)} \\ & + (\mathbf{t}_{\mathbf{z}^{(0)}} - \hat{\mathbf{t}}_{\mathbf{s_1 z}^{(0)}})\mathbf{A}^{-1}\mathbf{Z}_{+}^{(\mathbf{0})}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{E}_{+} + (\hat{\mathbf{t}}_{\mathbf{s_1 z}} - \hat{\mathbf{t}}_{\mathbf{s_2 z}})\mathcal{A}^{-1}\mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\ddot{\mathbf{E}}_{+} \qquad (150) \\ \approx\ & \mathbf{t}_{\mathbf{z}^{(0)}}\boldsymbol{\beta} + \sum_{x\in s_1}\frac{E^{(D)}(x)}{\pi_1(x)} + \sum_{x\in s_2}\frac{\ddot{E}^{(D)}(x)}{\pi^{*}(x)} - \sum_{x\in s_1}\frac{\ddot{E}^{(D)}(x)}{\pi_1(x)} \\ & + \boldsymbol{\Delta}_{\mathbf{z}^{(0)}}\mathbf{A}^{-1}\mathbf{Z}_{+}^{(\mathbf{0})}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\mathbf{E}_{+} + \boldsymbol{\Delta}_{\mathbf{z}}\mathcal{A}^{-1}\mathbf{Z}_{+}\boldsymbol{\Sigma}_{+}\boldsymbol{\Pi}_{+}\ddot{\mathbf{E}}_{+} \qquad (151) \\ \approx\ & \mathbf{t}_{\mathbf{z}^{(0)}}\boldsymbol{\beta} + \sum_{x\in s_1}\frac{\phi(x)}{\pi_1(x)} + \sum_{x\in s_2}\frac{\omega(x)}{\pi^{(*)}(x)} - \sum_{x\in s_1}\frac{\ddot{E}^{(D)}(x)}{\pi_1(x)}, \qquad (152)\end{aligned}$$

where

$$\phi(x) = e^{(D)}(x) + \boldsymbol{\Delta}_{\mathbf{z}^{(0)}}\tilde{\mathbf{G}}_{\beta_{+}^{(0)}}(x)e^{(D_{+})}(x) \tag{153}$$

and

$$\omega(x) = \ddot{e}^{(D)}(x) + \boldsymbol{\Delta}_{\mathbf{z}}\tilde{\mathbf{G}}_{\beta_{+}}(x)\ddot{e}^{(D_{+})}(x) \tag{154}$$

with their terms defined in Section 2.2.

Referring to the formula for the variance of any two-phase estimator (115), the conditional expectation of $\hat{t}_{y,2pm}$ is

$$\mathbb{E}_{2|1}\left[\hat{t}_{y,2pm}\right] = \mathbf{t}_{\mathbf{z}^{(0)}}\boldsymbol{\beta} + \sum_{x\in s_1}\frac{\phi(x)}{\pi_1(x)} \tag{155}$$

because

- the conditional expectation of the second and third term in (150) are equal to the third term, so they cancel each other
- the conditional expectation of the fifth term is also zero, because the conditional expectation of $\boldsymbol{\Delta}_{\mathbf{z}}$ is zero

The true total of true prediction $\mathbf{t}_{\mathbf{z}^{(\mathbf{0})}}\boldsymbol{\beta}$ does not contribute to the variance because it is a constant.

The conditional variance of $\hat{t}_{y,2pm}$ is specified by

$$\mathbb{V}_{2|1}\left[\hat{t}_{y,2pm}\right] = \sum_{x \in s_2} \frac{\omega(x)}{\pi^{(*)}(x)} \tag{156}$$

because all the other terms in (152) are constants, assuming a fixed sample of the first phase.

From this point on, the derivation of the variance estimator is a direct analogy of the derivations made in the previous section devoted to variance of the two-phase regression estimator of the total without exhaustive auxiliaries. It is obvious that the structure of the variance estimator is identical to (138), and that only $y^{(D)}(x)$ is replaced by $\phi(x)$ and $\rho(x)$ is replaced by $\omega(x)$. So, the final variance is (19) as presented in Section 2.2.

B Derivation of variances of two-phase ratio estimators

In the following four sections, the conditional expectations and covariances are derived for the specific combinations of two-phase generalised estimators in the ratio. The derivations assume identical samples used by the total estimators in the numerator and denominator. The terms γ in the variance estimator (42) are determined by these quantities only. The complete derivation of the variance estimator, including the specific definition of γ terms, is not given for space reasons. The main steps and logic combine the approximation of the ratio, as presented in Section 3 and the derivation of the two-phase variance included in Annex A.2.

B.1 Numerator and denominator are two-phase estimators with auxiliaries in the first-phase sample

This type of ratio is defined according to formula (25) with numerator and denominator corresponding to $\hat{t}_{y,2p}$ according to (2). Referring to (25), we can write

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2p} = \mathbf{I_D \Pi_D Y'_{1,D}} + \Delta'_{1,z^{(1)}}\mathbf{G}_{\mathbf{1},\beta^{(1)}_+}\mathbf{\Pi}_+\mathbf{Y}'_{\mathbf{1},+} - \\ - \hat{r}_{1,2}\left(\mathbf{I_{2,D}\Pi_D Y'_{2,D}} + \Delta'_{2,z^{(1)}}\mathbf{G}_{\mathbf{2},\beta^{(1)}_+}\mathbf{\Pi}_+\mathbf{Y}'_{\mathbf{2},+}\right),\end{aligned} \tag{157}$$

that can be detailed as follows

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2p} = \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi^*(x)} + \\ + \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi_1(x)} - \\ - \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\dot{e}_2^{(D)(x)}}{\pi^*(x)},\end{aligned} \tag{158}$$

and simplified to

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2p} = \sum_{x\in s_1}\frac{y_1^{(D)} - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi_1(x)} + \sum_{x\in s_2}\frac{\dot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi^*(x)} - \\ - \sum_{x\in s_1}\frac{\dot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi_1(x)}.\end{aligned} \tag{159}$$

For the conditional expectation over the second phase given the first-phase sample we have

$$\begin{aligned}\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right] = \mathbb{E}_{2|1}\left[\sum_{x\in s_1}\frac{y_1^{(D)} - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi_1(x)}\right] + \mathbb{E}_{2|1}\left[\sum_{x\in s_2}\frac{\dot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi^*(x)}\right] \\ - \mathbb{E}_{2|1}\left[\sum_{x\in s_1}\frac{\dot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi_1(x)}\right]\end{aligned} \tag{160}$$

The conditional expectation of the first term is just the term itself, because it is the estimate using the first-phase sample that is fixed by the expectation. The second term

is an unbiased estimator of the third term, so its conditional expectation equals the third term. Finally, the expectation of the third term is the term itself because it is a sum over the first-phase sample, which is fixed by the expectation. So, we obtain the following result:

$$\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x \in s_1} \frac{y_1^{(D)}(x) - \hat{r}_{1,2} y_2^{(D)}(x)}{\pi_1(x)}. \tag{161}$$

The conditional variance (over the second phase given the first-phase sample) is zero for the first and third term of (159) because these are fixed if the first-phase sample is fixed too. So, the conditional variance of $\hat{r}_{1,2}$ involves the second term only:

$$\mathbb{V}_{2|1}\left[\hat{r}_{1,2}\right] \approx \mathbb{V}_{2|1}\left(\sum_{x \in s_2} \frac{\dot{e}_1^{(D)}(x) - \hat{r}_{1,2} \dot{e}_2^{(D)}(x)}{\pi^*(x)}\right) \tag{162}$$

Adding the estimator of variance of (161) to the expectation of estimated variance (162) over the first phase leads to the variance estimator (42), where $\gamma_1(x) = y_1^{(D)}(x) - \hat{r}_{1,2} y_2^{(D)}(x)$, and $\gamma_2(x) = \rho_1(x) + \hat{r}_{1,2}\rho_2(x)$, see (5).

B.2 Numerator is a two-phase estimator with first-phase auxiliaries, denominator is a two-phase estimator with first-phase auxiliaries and exhaustive auxiliaries

This estimator uses $\hat{t}_{y,2p}$ according to (2) as $\hat{t}_1$ and $\hat{t}_{y,2pm}$ according to (16) as $\hat{t}_2$ of the ratio (23). Substituting to (25) we obtain:

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} &= \mathbf{I_{1,D}\Pi_D Y'_{1,D}} + \Delta'_{1,z^{(1)}}\mathbf{G}_{\mathbf{1},\beta^{(1)}_+}\mathbf{\Pi}_+\mathbf{Y'_{1},_+} \\ &- \hat{r}_{1,2}\left\{\mathbf{I_D\Pi_D Y'_{2,D}} + \Delta'_{2,z}\mathbf{G_{2},_{\beta_+}\Pi}_+\mathbf{Y'_{2},_+} + \Delta'_{2,z^{(0)}}\mathbf{G}_{\mathbf{2},\beta^{(0)}_+}\mathbf{\Pi}_+\mathbf{Y'_{2},_+}\right\}\end{aligned} \tag{163}$$

and further

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} &= \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi^*(x)} - \hat{r}_{1,2}t_{\tilde{y}_2} \\ &+ \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\ddot{e}_2^{(D)}(x) + \hat{r}_{1,2}\tilde{y}_2(x)}{\pi_1(x)} \\ &- \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)},\end{aligned} \tag{164}$$

where $\tilde{y}_2(x) = y_2^{(D)}(x) - e_2^{(D)}(x)$ is the prediction of $y_2(x)$ using the model with exhaustive auxiliaries and $t_{\tilde{y}_2}$ is the true sum of these predictions in D (region of estimation). After simplification we obtain:

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2p} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} &= \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \hat{r}_{1,2}e_2^{(D)}(x)}{\pi_1(x)} - \hat{r}_{1,2}t_{\tilde{y}_2} \\ &+ \sum_{x\in s_2}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}^{(2cp)}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)} - \sum_{x\in s_1}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\ddot{e}_2^{(D)}(x)}{\pi^*(x)},\end{aligned} \tag{165}$$

The conditional expectation of (165) given the first-phase follows

$$\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \hat{r}^{(2cp)}_{1,2}e_2^{(D)}(x)}{\pi_1(x)} - \hat{r}_{1,2}t_{\tilde{y}_2}. \tag{166}$$

Note that the last term is a constant, so it vanishes from the variance of $\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right]$ over the first phase. Concerning the conditional variance of (165), all terms except the second vanish because they do not vary with the first phase sample, so we can write

$$\mathbb{V}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x\in s_2}\frac{\dot{e}_1^{(D)}(x) - \hat{r}^{(2cp)}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)}. \tag{167}$$

The final estimator of variance corresponds to (42) with $\gamma_1(x) = y_1(x) - \hat{r}_{1,2}\phi_2(x)$ and $\gamma_2(x) = \rho_1(x) - \hat{r}_{1,2}\omega_2(x)$, see (20) and (21).

B.3 Numerator is a two-phase estimator with the first-phase and exhaustive auxiliaries, denominator is a two-phase estimator with the first-phase auxiliaries

This estimator uses $\hat{t}_{y,2pm}$ according to (16) as $\hat{t}_1$ and $\hat{t}_{y,2p}$ according to (2) as $\hat{t}_2$ of the ratio (23). Substituting to (25) we obtain:

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2p} &= \mathbf{I_D \Pi_D Y'_{1,D}} + \Delta'_{1,z}\mathbf{G}_{\mathbf{1},\beta_+}\mathbf{\Pi}_+\mathbf{Y}'_{\mathbf{1},+} + \Delta'_{1,z^{(0)}}\mathbf{G}_{\mathbf{1},\beta_+^{(\mathbf{0})}}\mathbf{\Pi}_+\mathbf{Y}'_{\mathbf{1},+} \\ &- \hat{r}_{1,2}\left\{\mathbf{I_{2,D}\Pi_D Y'_{2,D}} + \Delta'_{2,z^{(1)}}\mathbf{G}_{\mathbf{2},\beta_+^{(\mathbf{1})}}\mathbf{\Pi}_+\mathbf{Y}'_{\mathbf{2},+}\right\}\end{aligned} \tag{168}$$

and further

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2p} &= \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi^*(x)} + t_{\tilde{y}_1} \\ &+ \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\dot{e}_2^{(D)}(x) - y_1^{(D)}(x) + e_1^{(D)}(x)}{\pi_1(x)} - \\ &- \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\dot{e}_2^{(D)}}{\pi^*(x)},\end{aligned} \tag{169}$$

and simplify to obtain the final form:

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} &= \sum_{x\in s_1}\frac{e_1^{(D)}(x) - \hat{r}_{1,2}^{(2cp)}y_2^{(D)}(x)}{\pi_1(x)} + t_{\tilde{y}_1} + \\ &+ \sum_{x\in s_2}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}^{(2cp)}\dot{e}_2^{(D)}}{\pi^*(x)} - \sum_{x\in s_1}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}(x)}{\pi^*(x)},\end{aligned} \tag{170}$$

The conditional expectation of (170) given the first-phase sample is as it follows:

$$\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x\in s_1}\frac{e_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi_1(x)} + t_{\tilde{y}_1}. \tag{171}$$

Note that the last term $t_{\tilde{y}_1}$ is a constant, so it vanishes from the variance of $\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right]$ over the first phase. Concerning the conditional variance of (170), all terms except the second vanish because they do not vary given the first phase sample, so we can write

$$\mathbb{V}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x\in s_2}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\dot{e}_2^{(D)}}{\pi^*(x)}. \tag{172}$$

The final estimator of variance corresponds to (42) with $\gamma_1(x) = \phi_1(x) - \hat{r}_{1,2}y_2(x)$ and $\gamma_2(x) = \omega_1(x) - \hat{r}_{1,2}\rho_2(x)$, see (20), (21) and (5).

B.4 Numerator and denominator are two-phase estimators with the first-phase and exhaustive auxiliaries

This estimator uses $\hat{t}_{y,2pm}$ according to the formula (16) as $\hat{t}_1$ and $\hat{t}_2$ of the ratio (23). When it comes to the derivation of variance, we again start from the approximation (25) to which we substitute (16):

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} = \mathbf{I_D}\mathbf{\Pi_D}\mathbf{Y'_{1,D}} + \Delta'_{1,z}\mathbf{G}_{\mathbf{1},\beta_+}\mathbf{\Pi}_+\mathbf{Y'}_{\mathbf{1},+} + \Delta'_{1,z^{(0)}}\mathbf{G}_{\mathbf{1},\beta^{(\mathbf{0})}_+}\mathbf{\Pi}_+\mathbf{Y'}_{\mathbf{1},+} \\ - \hat{r}_{1,2}\left\{\mathbf{I}_{\mathbf{2,D}}\mathbf{\Pi_D}\mathbf{Y'_{2,D}} + \Delta'_{2,z}\mathbf{G}_{\mathbf{2},\beta_+}\mathbf{\Pi}_+\mathbf{Y}_{\mathbf{2},+} + \Delta'_{2,z^{(0)}}\mathbf{G}_{\mathbf{2},\beta^{(\mathbf{0})}_+}\mathbf{\Pi}_+\mathbf{Y'}_{\mathbf{2},+}\right\}\end{aligned} \tag{173}$$

Next we change the matrix form for sums

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} = \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x)}{\pi^*(x)} + t_{\tilde{y}_1} - \hat{r}^{(2cp)}_{1,2}t_{\tilde{y}_2} + \\ + \sum_{x\in s_1}\frac{y_1^{(D)}(x) - \ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\ddot{e}_2^{(D)}(x)}{\pi_1(x)} - \\ - \sum_{x\in s_1}\frac{y_1^{(D)}(x) - e_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}e_2^{(D)}(x)}{\pi_1(x)} - \\ - \sum_{x\in s_2}\frac{y_1^{(D)}(x) - \ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}y_2^{(D)}(x) + \hat{r}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)},\end{aligned} \tag{174}$$

and simplify to obtain the final form:

$$\begin{aligned}\hat{t}^{(1)}_{y_1,2pm} - \hat{r}_{1,2}\hat{t}^{(2)}_{y_2,2pm} = \sum_{x\in s_1}\frac{e_1^{(D)}(x) - \hat{r}_{1,2}e_2^{(D)}(x)}{\pi_1(x)} + t_{\tilde{y}_1} - \hat{r}_{1,2}t_{\tilde{y}_2} + \\ + \sum_{x\in s_2}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)} - \sum_{x\in s_1}\frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\ddot{e}_2^{(D)}(x)}{\pi^*(x)},\end{aligned} \tag{175}$$

where

- $t_{\tilde{y}_1} = t'_{z^{(0)},1}\tilde{\mathbf{G}}^{(0)}_{\beta_+,1}\mathbf{\Pi}_+\mathbf{Y}'_{+,1}$ is the total of model predictions of the variable y_1 in the estimation cell D. It is calculated as a product of a known total $t_{z^{(0)},1}$ of the auxiliary vector (known anywhere in D) and parameters of the corresponding model.
- $t_{\tilde{y}_2} = t'_{z^{(0)},2}\tilde{\mathbf{G}}^{(0)}_{\beta_+,2}\mathbf{\Pi}_+\mathbf{Y}'_{+,2}$ is total of model predictions of the variable y_2.

The conditional expectation of (175) given the first-phase sample is given by the formula

$$\mathbb{E}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x\in s_1}\frac{e_1^{(D)}(x) - \hat{r}_{1,2}e_2^{(D)}(x)}{\pi_1(x)} + t_{\tilde{y}_1} - \hat{r}_{1,2}t_{\tilde{y}_2}. \tag{176}$$

Note that the last two terms are constants, so they vanish from the variance $\mathbb{E}_{2|1}\left(\hat{r}_{1,2}\right)$ over the first phase. Concerning the conditional variance of (175), all terms except the

fourth vanish because they do not vary given the first phase sample, so we can write

$$\mathbb{V}_{2|1}\left[\hat{r}_{1,2}\right] \approx \sum_{x \in s_2} \frac{\ddot{e}_1^{(D)}(x) - \hat{r}_{1,2}\ddot{e}_2^{(D)}}{\pi^*(x)}. \tag{177}$$

The final estimator of variance of $\hat{r}_{1,2}$ again matches (42) with $\gamma_1(x) = \phi_1(x) - \hat{r}_{1,2}\phi_2(x)$ and $\gamma_2(x) = \omega_1(x) - \hat{r}_{1,2}\omega_2(x)$, see (20) and (21).

C Derivation of covariances of two-phase total estimators

The covariance of two two-phase total estimators $\hat{t}_{y_1,2p}$ and $\hat{t}_{y_2,2p}$ is defined

$$\mathbb{C}\left[\hat{t}_{y_1,2p}, \hat{t}_{y_2,2p}\right] = \mathbb{C}_1\left[\mathbb{E}_{2|1}(\hat{t}_{y_{y_1,2p}}), \mathbb{E}_{2|1}(\hat{t}_{y_2,2p})\right] + \mathbb{E}_1\left[\mathbb{C}_{2|1}\left(\hat{t}_{y_1,2p}, \hat{t}_{y_2,2p}\right)\right]. \tag{178}$$

It is a sum of the covariance of the conditional expectations given the first-phase samples and the expectation over the first-phase of the second-phase covariance given the first-phase samples, see Särndal *et al.* [2003, p. 136]. In this annex and its sections, the covariances are derived for totals that do not necessarily use identical samples. If the samples are disjoint, the covariances are zero. If the samples are identical, the formulas are still correct, but could be simplified. The objective was primarily to determine covariances of the totals using unequal samples with partial intersection, that is, samples with a definite number of points shared between the two total estimators. This situation is quite common if estimates use several panels and correspond to the so-called moving window averages. The covariances are then needed to estimate variance of difference of such moving window estimators, totals, and ratios.

In next section the covariance of single-phase and two-phase total is derived in full detail. It is quite obvious that covariance of two totals from which one or both are two-phase can be expressed by

$$\begin{aligned}
\hat{\mathbb{C}}\left[\hat{t}_{y_1,2p}, \hat{t}_{y_2,2p}\right] = & \sum_{x \in s_2^{(m)}} \frac{\tau_1(x)\tau_2(x)}{\pi_{2|1}^{(m)}(x)\pi_1^{(1)}(x)\pi_1^{(2)}(x)} + \\
& + \sum_{x \in s_2^{(1)}} \sum_{\substack{x' \in s_2^{(2)} \\ x' \neq x}} \frac{\tau_1(x)\tau_2(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')} \times \frac{\pi_1^{(1,2)}(x,x') - \pi_1^{(1)}(x)\pi_1^{(2)}(x')}{\pi^{*(1,2)}(x,x')} + \\
& + \sum_{x \in s_2^{(m)}} \frac{\tau_3(x)\tau_4(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)} \left[1 - \pi_{2|1}^{(m)}\right] + \\
& + \sum_{x \in s_2^{(1)}} \sum_{\substack{x' \in s_2^{(2)} \\ x' \neq x}} \frac{\tau_3(x)\tau_4(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')} \times \frac{\pi_{2|1}^{(1,2)}(x,x') - \pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')}{\pi_{2|1}^{(1,2)}(x,x')}.
\end{aligned} \tag{179}$$

where the terms $\tau_1(x)$, $\tau_2(x)$, $\tau_3(x)$ and $\tau_4(x)$ depend on the combination of estimator types (single-phase or two-phase, with or without exhaustive auxiliaries).

Conditional expectations and covariances corresponding to the first and second terms of (178) are the only inputs on which $\tau(x)$ terms depend. Conditional expectations (given the first-phase sample) correspond to $\tau_1(x)$ and $\tau_2(x)$. The terms of the conditional covariance correspond to $\tau_3(x)$ and $\tau_4(x)$. There is no dependence on the order of estimator types and the $\tau(x)$ terms because the covariance depends on the combination of estimator types, not on their order.

Annexes C.1, C.2 and C.3 present conditional expectations for all estimator types considered. Similarly, the terms within conditional covariances are derived for all estimator types.

C.1 Covariance of single-phase total estimator with no auxiliaries and two-phase total estimator with first-phase auxiliaries

The single-phase estimator is defined

$$\hat{t}_{y_1,1p} = \sum_{x \in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{(1)}(x)} \tag{180}$$

Assume the sample $s_2^{(1)}$ was obtained in a single phase. Even if the sample was obtained in more phases, the estimator remains unbiased as long as the inclusion density $\pi^{(1)}(x)$ corresponds to $\pi^{*(1)}(x)$ defined in Section 2.1.

For the evaluation of covariance with two-phase estimators, it is convenient and possible to rewrite the single-phase estimator as a two-phase regression estimator

$$\hat{t}_{y_1,1p} = \sum_{x \in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{*(1)}(x)} + \sum_{x \in s_1^{(1)}} \frac{\left[y_1^{(D)}(x) - \dot{e}_1^{(D)}(x)\right]}{\pi_1^{(1)}(x)} - \sum_{x \in s_2^{(1)}} \frac{\left[y_1^{(D)}(x) - \dot{e}_1^{(D)}(x)\right]}{\pi^{*(1)}(x)}, \tag{181}$$

where the sampling densities $\pi^{(1)}(x)$ $\pi^{*(1)}(x)$ and $\pi_1^{(1)}(x)$ are identical because all points in the first phase are selected in the second phase, i.e., the conditional sampling probability $\pi_{2|1}^{(1)}(x) = 1$. The two sums run over the same set of points because $s_1^{(1)} = s_2^{(2)}$, so they cancel each other.

The two-phase generalised regression estimator without exhaustive auxiliaries according to (2) can be written as follows

$$\hat{t}_{y_2,2p} = \sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x)}{\pi^{*(2)}(x)} + \sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x) - \dot{e}_2^{(D)}(x)}{\pi_1^{(2)}(x)} - \sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x) - \dot{e}_2^{(D)}(x)}{\pi^{*(2)}(x)}. \tag{182}$$

Referring to the first term of (178), the conditional expectations is defined

$$\begin{aligned} \mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right] = \mathbb{E}_{2|1}\left[\sum_{x \in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{*(1)}(x)}\right] + \mathbb{E}_{2|1}\left[\sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x)}{\pi_1^{(1)}(x)}\right] - \\ - \mathbb{E}_{2|1}\left[\sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x) - \dot{e}_2^{(D)}(x)}{\pi^{*(2)}(x)}\right] \end{aligned} \tag{183}$$

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right] = \sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)} + \sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x)}{\pi_1^{(1)}(x)} - \sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x) - \dot{e}_1^{(D)}(x)}{\pi_1^{(1)}(x)} \tag{184}$$

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right] = \sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}, \tag{185}$$

and

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2p}\right] = \sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)} \tag{186}$$

According to the general formula for covariance as an expectation of a product minus the product of expectations, we have

$$\begin{aligned}
&\mathbb{C}_1\left\{\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right],\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2p}\right]\right\}=\mathbb{C}_1\left\{\sum_{x\in s_1^{(1)}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)},\sum_{x\in s_1^{(2)}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\right\}\\
&=\mathbb{C}_1\left\{\sum_{x\in s_1^{(1)}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)},\sum_{x\in s_1^{(2)}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\right\}=\mathbb{E}_1\left\{\sum_{x\in s_1^{(1)}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}\sum_{x\in s_1^{(2)}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\right\}-\\
&-\mathbb{E}_1\left\{\sum_{x\in s_1^{(1)}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}\right\}\mathbb{E}_1\left\{\sum_{x\in s_1^{(2)}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\right\}
\end{aligned}\tag{187}$$

For the expectation of the product, we can write

$$\begin{aligned}
&\mathbb{E}_1\left\{\sum_{x\in s_1^{(1)}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}\sum_{x\in s_1^{(2)}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\right\}=\mathbb{E}_1\left\{\sum_{x\in s_1^{(1)}}\sum_{x'\in s_1^{(2)}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}\right\}\\
&=\mathbb{E}_1\left\{\sum_{x\in s_1^{(m)}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}\right\}+\mathbb{E}_1\left\{\sum_{x\in s_1^{(1)}}\sum_{\substack{x'\in s_1^{(2)}\\x'\neq x}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\right\},
\end{aligned}\tag{188}$$

where $s_1^{(m)}=s_1^{(1)}\cap s_1^{(2)}$ is the matched sample of the first-phase defined in an analogy to (36).

The first expectation term of (188) equals

$$\begin{aligned}
&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1^{(m)}}\frac{y_1^{(D)}(x_i)y_2^{(D)}(x_i)}{\pi_1^{(1)}(x_i)\pi_1^{(2)}(x_i)}\right\}f_1^{(m)}\left[x_1,x_2,\ldots x_{n_1^{(m)}}\right]dx_1dx_2\ldots dx_{n_1^{(m)}}\\
&=\sum_{i=1}^{n_1^{(m)}}\int_{\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}f_{1,i}^{(m)}(x)dx=\int_{\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}\sum_{i=1}^{n_1^{(m)}}f_{1,i}^{(m)}(x)dx\\
&=\int_{\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}\pi_1^{(m)}(x)dx,
\end{aligned}\tag{189}$$

where the joint probability density $f_1^{(m)}[x_1,x_2,\ldots x_{n_1^{(m)}}]$ refers to simultaneous selection of all sample points of $s_1^{(m)}$ in the first phase, and the probability densities $f_{1,i}^{(m)}(x)$ refer to the selection of point x in the i-th draw. The region of integration $\mathcal{U}$ is the geographical space in which sampling is implemented (sampling frame) and the inclusion densities and pairwise inclusion densities (the sampling design) are defined. The transition from multiple to single integral is possible according to Fubini's theorem, which allows for iterative integration over each $x\in s_1^{(m)}$ in every step, so the number of points in the joint probability density is reduced by one in each iteration until only one point remains. The

final operation uses the definition of inclusion density in the matched sample

$$\pi_1^{(m)}(x) = \sum_{i=1}^{n_1^{(m)}} f_{1,i}^{(m)}(x). \tag{190}$$

The second expectation term for distinct x and x' is defined by

$$\begin{aligned}
&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1^{(1)}}\sum_{\substack{x_j\in s_1^{(2)}\\ x_j\neq x_i}}\frac{y_1^{(D)}(x_i)y_2^{(D)}(x_j)}{\pi_1^{(1)}(x_i)\pi_1^{(2)}(x_j)}\right\} f_1^{(1,2)}[x_1, x_2, \ldots x_{n_1^{(1)}} \\
&\quad \ldots x_{n_1^{(1)}+1}, x_{n_1^{(1)}+2}\ldots x_{n_1^{(1)}+n_1^{(2)}-1}]dx_1dx_2\ldots dx_{n_1^{(1)}}\ldots dx_{n_1^{(1)}+n_1^{(2)}-1} \\
&= \sum_{x_i\in s_1^{(1)}}\sum_{\substack{x_l\in s_1^{(2)}\\ x_l\neq x_i}}\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')} f_{1,i,l}^{(1,2)}\left[x_i\in s_1^{(1)}, x_l\in s_1^{(2)}\setminus\{x_i\}\right]dxdx' \\
&= \int\int_{\mathcal{U}\times\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\sum_{x_i\in s_1^{(1)}}\sum_{\substack{x_l\in s_1^{(2)}\\ x_l\neq x_i}} f_{1,i,l}^{(1,2)}\left[x_i\in s_1^{(1)}, x_l\in s_1^{(2)}\setminus\{x_i\}\right]dxx' \\
&= \int\int_{\mathcal{U}\times\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\pi_1^{(1,2)}(x,x')dxdx',
\end{aligned} \tag{191}$$

where Fubini's theorem is used to gradually integrate out all the points from the joint probability density until we get $f_{1,i,l}^{(1,2)}\left[x_i\in s_1^{(1)}, x_l\in s_1^{(2)}\setminus\{x_i\}\right]$, the probability density of simultaneous inclusion of distinct points $x = x_i$ as the i-th point in sample $s_1^{(1)}$, and $x' = x_l$ as the l-th point in the sample $s_1^{(2)}$ with the point x removed. Note that it is assumed that samples $s_1^{(1)}$ and $s_1^{(2)}$ share at least one sample point, that is $s_1^{(m)} \neq \emptyset$.

The pairwise inclusion density $\pi_1^{(1,2)}(x,x')$ refers to the simultaneous inclusion of distinct points x and x' in the corresponding samples $s_1^{(1)}$ and $s_1^{(2)}$. It is defined by equation

$$\pi_1^{(1,2)}(x,x') = \pi_1^{(1,2)}(x_i,x_l) = \sum_{x_i\in s_1^{(1)}}\sum_{\substack{x_l\in s_1^{(2)}\\ x_l\neq x_i}} f_{1,i,l}\left[x_i\in s_1^{(1)}, x_l\in s_1^{(2)}\setminus\{x_i\}\right]. \tag{192}$$

Now, evaluate the product of the conditional expectations in (187)

$$\begin{aligned}
&\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\sum_{x_i\in s_1^{(1)}}\left\{\frac{y_1^{(D)}(x_i)}{\pi_1^{(1)}(x_i)}\right\}f_1^{(1)}(x_1,x_2,\ldots x_{n_1^{(1)}})dx_1dx_2\ldots dx_{n_1^{(1)}}\times\\
&\quad\times\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\sum_{x_j\in s_1^{(2)}}\left\{\frac{y_2^{(D)}(x_j)}{\pi_1^{(2)}(x_j)}\right\}f_1^{(2)}(x_1,x_2,\ldots x_{n_1^{(2)}})dx_1dx_2\ldots dx_{n_1^{(2)}}\\
&=\sum_{i=1}^{n_1^{(1)}}\int_{\mathcal{U}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}f_{1,i}^{(1)}(x)dx\sum_{i=1}^{n_1^{(2)}}\int_{\mathcal{U}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}f_{1,i}^{(2)}(x)dx\\
&=\int_{\mathcal{U}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}\sum_{i=1}^{n_1^{(1)}}f_{1,i}^{(1)}(x)dx\int_{\mathcal{U}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\sum_{i=1}^{n_1^{(2)}}f_{1,i}^{(2)}(x)dx\\
&=\int_{\mathcal{U}}\frac{y_1^{(D)}(x)}{\pi_1^{(1)}(x)}\pi_1^{(1)}(x)dx\int_{\mathcal{U}}\frac{y_2^{(D)}(x)}{\pi_1^{(2)}(x)}\pi_1^{(2)}(x)dx\\
&=\int_{\mathcal{U}}y_1^{(D)}(x)dx\int_{\mathcal{U}}y_2^{(D)}(x)dx=\int\int_{\mathcal{U}\times\mathcal{U}}y_1^{(D)}(x)y_2^{(D)}(x')dxdx'
\end{aligned}\tag{193}$$

The final covariance of the conditional expectations over the first phase combines the previous two results

$$\begin{aligned}
&\mathbb{C}_1\left\{\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right],\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2p}\right]\right\}=\int_{\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_1^{(1)}(x)\pi_1^{(2)}(x)}\pi_1^{(m)}(x)dx+\\
&+\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\left[\pi_1^{(1,2)}(x,x')-\pi_1^{(1)}(x)\pi_1^{(2)}(x')\right]dxdx'
\end{aligned}\tag{194}$$

Its estimator using second-phase sample points takes the following form

$$\begin{aligned}
&\hat{\mathbb{C}}_1\left\{\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1p}\right],\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2p}\right]\right\}=\sum_{x\in s_2^{(m)}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_{2|1}^{(m)}(x)\pi_1^{(1)}(x)\pi_1^{(2)}(x)}+\\
&+\sum_{x\in s_2^{(1)}}\sum_{\substack{x'\in s_2^{(2)}\\x'\neq x}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\times\frac{\pi_1^{(1,2)}(x,x')-\pi_1^{(1)}(x)\pi_1^{(2)}(x')}{\pi^{*(1,2)}(x,x')}.
\end{aligned}\tag{195}$$

The pairwise inclusion density over the two phases in both samples is defined

$$\pi^{*(1,2)}(x,x')=\pi_1^{(1,2)}(x,x')\pi_{2|1}^{(1,2)}(x,x'),\tag{196}$$

where

$$\pi_{2|1}^{(1,2)}(x,x')\;=\;\mathbb{P}\left[x\in s_2^{(1)},\;x'\in s_2^{(2)}\;\middle|\;x\in s_1^{(1)},\;x'\in s_1^{(2)}\right]\tag{197}$$

is the probability of including point x in $s_2^{(1)}$ given its membership to $s_1^{(1)}$ and at the same time including x' in $s_2^{(2)}$ given its membership to $s_1^{(2)}$.

The term $\pi_{2|1}^{(m)}(x)$ is the conditional probability of selecting point x from the first phase matched sample $s_1^{(m)}$ to the second phase matched sample $s_2^{(m)}$

$$\pi_{2|1}^{(m)}(x)=\mathbb{P}\left[x\in s_2^{(m)}|\;x\in s_1^{(m)}\right]\tag{198}$$

The conditional covariance within the second term of (178) is

$$\mathbb{C}_{2|1}\left[\hat{t}_{y_{1,1p}}, \hat{t}_{y_{2,2p}}\right] = \mathbb{C}_{2|1}\left[\sum_{x\in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{*(1)}(x)}, \sum_{x\in s_2^{(2)}} \frac{\dot{e}_2^{(D)}(x)}{\pi^{*(2)}(x)}\right], \tag{199}$$

where the first term corresponds to the single-phase estimator, and the second term is the only variable part remaining from (182) if the first phase sample is kept constant, i.e., if it does not vary.

This conditional covariance refers to the finite population of the first phase points, so it is defined

$$\begin{aligned}\mathbb{C}_{2|1}&\left[\hat{t}_{y_{1,1p}}, \hat{t}_{y_{2,2p}}\right] = \\ &= \sum_{x\in s_1^{(1)}} \sum_{x'\in s_1^{(2)}} \frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')}\left[\pi_{2|1}^{(1,2)}(x,x') - \pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')\right]\end{aligned} \tag{200}$$

Below, single and pairwise conditional probabilities are derived for a specific but practically relevant situations. Imagine that we have three independently generated first-phase samples, that is, that we have three panels s_{11}, s_{12} and s_{13}. From these three second-phase samples s_{21}, s_{22} and s_{23} are drawn, with inclusion probabilities $\pi_{21}(x)$, $\pi_{22}(x)$ and $\pi_{23}(x)$. The first-phase sample $s_1^{(1)}$ of the numerator is created by merging the samples s_{11} and s_{12}, so $s_1^{(1)} = s_{11}\cup s_{12}$. Similarly, the second first-phase sample is defined $s_1^{(2)} = s_{1,2}\cup s_{1,3}$. The second-phase samples are generated analogously, so $s_2^{(1)} = s_{2,1}\cup s_{2,2}$ and $s_2^{(2)} = s_{2,2}\cup s_{2,3}$, which means that they are not obtained by subsampling the merged first-phase samples $s_1^{(1)}$ and $s_1^{(2)}$.

Such first and second phase samples are part of the proposed information systems that use panel merging, see Section 6.4. The conditional probability of inclusion into the sample $s_2^{(1)}$ is defined

$$\pi_{2|1}^{(1)}(x) = \begin{cases} \pi_{21}(x) & \text{if } x \text{ belongs to } s_{11} \\ \pi_{22}(x) & \text{if } x \text{ belongs to } s_{12} \\ \pi_{21}(x) + \pi_{22}(x) - \pi_{21}(x)\pi_{22}(x) & \text{if } x \text{ belongs to } s_{11}\cap s_{12}, \end{cases} \tag{201}$$

and similarly for the sample $s_2^{(2)}$

$$\pi_{2|1}^{(2)}(x) = \begin{cases} \pi_{22}(x) & \text{if } x \text{ belongs to } s_{12} \\ \pi_{23}(x) & \text{if } x \text{ belongs to } s_{13} \\ \pi_{22}(x) + \pi_{23}(x) - \pi_{22}(x)\pi_{23}(x) & \text{if } x \text{ belongs to } s_{12}\cap s_{13}, \end{cases} \tag{202}$$

and for the matched sample $s_2^{(m)}$

$$\pi_{2|1}^{(m)}(x) = \pi_{22}(x). \tag{203}$$

Now, consider the set of indices $\mathcal{Q}$ of the first-phase panels unified into $s_1^{(1)}$, and analogous sets of indices $\mathcal{V}$ and $\mathcal{M}$ of panels unified into $s_1^{(2)}$ and $s_1^{(m)}$ respectively.

The pairwise conditional probability $\pi_{2|1}^{(1,2)}(x,x')$ is evaluated by the following equation

$$\pi_{2|1}^{(1,2)}(x,x') = \sum_{q\in\mathcal{Q}}\sum_{v\in\mathcal{V}} \pi_{2|1}^{(1,2)}\left[x\in s_{2q}, x'\in s_{2v}\right], \tag{204}$$

where

$$\pi_{2|1}^{(1,2)}(x,x') = \begin{cases} 1. & q \neq v & \pi_{2q}(x)\pi_{2v}(x') \\ 2. & q = v \wedge x \neq x' & \pi_{2q}(x,x') \\ 3. & q = v \wedge x = x' & \pi_{2q}(x). \end{cases} \tag{205}$$

The first case for $q \neq v$ applies for both situations $x = x'$ and $x \neq x'$. The second case is the pairwise probability of selecting two distinct points x and x' from the first-phase sample sample s_{1q} to the second-phase sample s_{2q}. The third case reduces to one possible situation, namely $q = v = 2$ and $\pi_{2|1}^{(1,2)}(x,x') = \pi_{22}$. The only way the point x arrives in $s_2^{(1)}$ and simultaneously in $s_2^{(2)}$ is that it is included in s_{22}.

Imagine, for instance, a point x belonging to s_{11} and also to s_{12} and a point $x' \neq x$ belonging to s_{13}. The first sum of (204) will have two terms and the second only one term corresponding to two and one original sample containing x and x'. The pairwise probability will be a sum of $\pi_{21}\pi_{23}$ and $\pi_{22}\pi_{23}$. This is logical because the samples s_{11} and s_{13} are independent, as are the samples s_{12} and s_{13}. Another example corresponding to the second case is when two distinct points x and x' belong to s_{12}. The conditional pairwise probability will be $\pi_{22}(x,x')$.

The above derivations for two distinct and one common sample is used for demonstration only. In practice, the samples $s_2^{(1)}$, $s_2^{(2)}$ and $s_2^{(m)}$ are often unions of more then just two panels. The formula (204) is generally usable. The generalisation for single inclusion probabilities is the following

$$\begin{gathered} \pi_{2|1}^{(1)}(x) = 1 - \prod_{q\in\mathcal{Q}} \left[1 - \pi_{2q}(x)\right] \qquad \pi_{2|1}^{(2)}(x) = 1 - \prod_{v\in\mathcal{V}} \left[1 - \pi_{2v}(x)\right] \\ \pi_{2|1}^{(m)}(x) = 1 - \prod_{m\in\mathcal{M}} \left[1 - \pi_{2m}(x)\right]. \end{gathered} \tag{206}$$

It might happen that a panel is included more than once in any of the three samples $s_2^{(m)}$, $s_2^{(2)}$ and $s_2^{(m)}$ (and also in the corresponding first-phase samples), each time with a different reference year (or set of reference years). In such cases, the values of target variable must be averaged over the reference years because only one observation would be preserved from each panel during the union operation.

The first-phase samples (panels) of NFIs are typically not fully independent in the sense that they are derived from one large sample (the whole sampling grid) by randomised assignment of survey years to plots (or their clusters, if used). Each sample point belongs to only one single panel of given sampling phase. The points $x \in s_2^{(1)}$ and $x' \in s_2^{(2)}$ are identical if only if they belong to the same first- and second-phase panel, which means they also belong to the matched sample $s_2^{(m)}$. Obviously, the following equality holds $\pi_{2|1}^{(1,2)}(x,x) = \pi_{2|1}^{(1)}(x) = \pi_{2|1}^{(2)}(x) = \pi_{2|1}^{(m)}(x)$. Finally, the conditional covariance (200) can be reformulated as follows

$$\begin{aligned} \mathbb{C}_{2|1}\left[\hat{t}_{y_{1,1p}}, \hat{t}_{y_{2,2p}}\right] &= \sum_{x\in s_1^{(m)}} \frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)} \left[1 - \pi_{2|1}^{(m)}(x)\right] \pi_{2|1}^{(m)}(x) + \\ &+ \sum_{x\in s_1^{(1)}} \sum_{x'\in s_1^{(2)}} \frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')} \left[\pi_{2|1}^{(1,2)}(x,x') - \pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')\right] \end{aligned} \tag{207}$$

Referring to (178), the expectation of the first term in (207) over the first phase samples is defined

$$\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x\in s_1^{(m)}}\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}(x)\right]\pi_{2|1}^{(m)}(x)\times f_1^{(m)}(x_1,x_2\ldots x_{n_1^{(m)}})\right\}$$

$$dx_1dx_2\ldots dx_{n_1^{(m)}}= \tag{208}$$

$$=\int_{\mathcal{U}}\left\{\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}(x)\right]\pi_{2|1}^{(m)}(x)\times\sum_{i=1}^{n_1^{(m)}}f_{1,i}^{(m)}(x)\right\}dx \tag{209}$$

$$=\int_{\mathcal{U}}\left\{\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}(x)\right]\pi_{2|1}^{(m)}(x)\times\pi_1^{(m)}(x)\right\}dx \tag{210}$$

$$=\int_{\mathcal{U}}\left\{\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}(x)\right]\times\pi_1^{*(m)}(x)\right\}dx \tag{211}$$

Its estimator is computed from the second-phase matched sample $s_2^{(m)}$ dividing by $\pi^{*(m)}(x)$

$$\sum_{x\in s_2^{(m)}}\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}\right] \tag{212}$$

The expectation of the double sum in (207) over the set of first phase samples is given by

$$\int_{\mathcal{U}}\int_{\mathcal{U}}\cdots\int_{\mathcal{U}}\left\{\sum_{x_i\in s_1^{(1)}}\sum_{\substack{x_j\in s_1^{(2)}\\x_j\neq x}}\frac{y_1^D(x_i)\dot{e}_2^D(x_j)}{\pi^{*(1)}(x_i)\pi^{*(2)}(x_j)}\left[\pi_{2|1}^{(1,2)}(x_i,x_j)-\pi_{2|1}^{(1)}(x_i)\pi_{2|1}^{(2)}(x_j)\right]\right\}\times \tag{213}$$

$$\times\, f_1^{(1,2)}(x_1,x_2\ldots x_{n_1^{(1)}+n_1^{(2)}-1})dx_1dx_2\ldots dx_{n_1^{(1)}+n_1^{(2)}-1}.$$

It can be simplified to

$$\int\int_{\mathcal{U}\times\mathcal{U}}\frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')}\left[\pi_{2|1}^{(1,2)}(x,x')-\pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')\right]\times \tag{214}$$

$$\times\;\pi_1^{(1,2)}(x,x')dxdx',$$

and estimated by

$$\sum_{x\in s_2^{(1)}}\sum_{\substack{x'\in s_2^{(2)}\\x'\neq x}}\frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')}\times\frac{\pi_{2|1}^{(1,2)}(x,x')-\pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')}{\pi_{2|1}^{(1,2)}(x,x')} \tag{215}$$

The estimator of the expectation of the conditional covariance (207) over the set of first phase samples is the sum of the above estimators (212) and (215) above

$$\widehat{\mathbb{E}}\left\{\mathbb{C}_{2|1}\left[\hat{t}_{y_{1,1p}},\hat{t}_{y_{2,2p}}\right]\right\}=\sum_{x\in s_2^{(m)}}\frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}\right]$$
$$+\sum_{x\in s_2^{(1)}}\sum_{\substack{x'\in s_2^{(2)}\\x'\neq x}}\frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')}\times\frac{\pi_{2|1}^{(1,2)}(x,x')-\pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')}{\pi_{2|1}^{(1,2)}(x,x')}. \tag{216}$$

By combination of (195) and (216) the estimator of the covariance of the single-phase and two-phase total is obtained

$$\hat{\mathbb{C}}\left[\hat{t}_{y_{1,2p}},\hat{t}_{y_{2,2p}}\right]=\sum_{x\in s_2^{(m)}}\frac{y_1^{(D)}(x)y_2^{(D)}(x)}{\pi_{2|1}^{(m)}(x)\pi_1^{(1)}(x)\pi_1^{(2)}(x)}+$$
$$+\sum_{x\in s_2^{(1)}}\sum_{\substack{x'\in s_2^{(2)}\\x'\neq x}}\frac{y_1^{(D)}(x)y_2^{(D)}(x')}{\pi_1^{(1)}(x)\pi_1^{(2)}(x')}\times\frac{\pi_1^{(1,2)}(x,x')-\pi_1^{(1)}(x)\pi_1^{(2)}(x')}{\pi^{*(1,2)}(x,x')}+$$
$$+\sum_{x\in s_2^{(m)}}\frac{y_1^D(x)\dot{e}_2^D(x)}{\pi^{*(1)}(x)\pi^{*(2)}(x)}\left[1-\pi_{2|1}^{(m)}\right]+ \tag{217}$$
$$+\sum_{x\in s_2^{(1)}}\sum_{\substack{x'\in s_2^{(2)}\\x'\neq x}}\frac{y_1^D(x)\dot{e}_2^D(x')}{\pi^{*(1)}(x)\pi^{*(2)}(x')}\times\frac{\pi_{2|1}^{(1,2)}(x,x')-\pi_{2|1}^{(1)}(x)\pi_{2|1}^{(2)}(x')}{\pi_{2|1}^{(1,2)}(x,x')}.$$

C.2 Covariance of single-phase total with no auxiliaries and two-phase total with the first-phase and exhaustive auxiliaries

One estimator corresponds to single-phase estimator $\hat{t}_{y_1,1p}$ according to (180), and the other to two-phase generalised regression estimator $\hat{t}_{y_2,2pm}$ according to (16). The conditional expectation of the single-phase estimator is

$$\mathbb{E}_{2|1}\hat{t}_{y_1,1p} = \sum_{x \in s_1^{(1)}} \frac{y_1^{(D)}(x)}{\pi_1(x)} \tag{218}$$

see the derivation of (185). For the two-phase estimator we have

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2pm}\right] = \mathbb{E}_{2|1}\left[\begin{array}{l} \displaystyle\sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x)}{\pi^{*(2)}(x)} + \\ \displaystyle\sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x) - \ddot{e}_2^{(D)}(x)}{\pi_1^{(2)}(x)} - \sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x) - \ddot{e}_2^{(D)}(x)}{\pi^{*(2)}(x)} + \\ t_{\tilde{y}_2} - \displaystyle\sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x) - e_2^{(D)}(x)}{\pi_1^{(2)}(x)} \end{array}\right] \tag{219}$$

where the term $t_{\tilde{y}_2}$ corresponds to the known total of working model predictions in D, but this term does not appear in the covariance and its $\tau(x)$ terms, because it is a constant. The conditional expectation of the second term in the middle line equals the first term, so they cancel each other out. The expectation of the first sum equals

$$\mathbb{E}_{2|1}\left[\sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x)}{\pi^{*(2)}(x)}\right] = \sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x)}{\pi_1(x)} \tag{220}$$

and it cancels with the corresponding part of the last sum. The result follows

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_2,2pm}\right] = t_{\tilde{y}_2} - \sum_{x \in s_1^{(2)}} \frac{e_2^{(D)}(x)}{\pi_1^{(2)}(x)}. \tag{221}$$

Concerning conditional covariance, the single-phase estimator is taken as a whole, and from the two-phase estimator shown within (219) only the sums over $s_2^{(2)}$ are kept, while the first cancels with the first part of the second one. Finally, only the second-phase estimator of the full model residuals remains.

$$\mathbb{C}_{1|2}\left[\hat{t}_{y_1,1p}, \hat{t}_{y_2,2pm}\right] = \mathbb{C}_{1|2}\left[\sum_{x \in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{*(1)}(x)}, \sum_{x \in s_2^{(2)}} \frac{\ddot{e}_2^{(D)}(x)}{\pi^{*(2)}(x)}\right] \tag{222}$$

C.3 Covariance of a single-phase total estimator with exhaustive auxiliaries and a two-phase total estimator with first-phase auxiliaries

The single-phase, generalised regression estimator is defined

$$\hat{t}_{y_1,1pm} = \mathbf{I_D \Pi_D Y'_{1,D}} + \mathbf{\Delta'_{1,z^{(0)}} G_{1,\beta_+^{(0)}} \Pi_+ Y'_{1,+}} \tag{223}$$

$$= \sum_{x \in s_2^{(1)}} \frac{y_1^{(D)}(x)}{\pi^{*(1)}(x)} + t_{\tilde{y}_1} - \sum_{x \in s_2} \frac{y_1^{(D)}(x) - e_1^{(D)}(x)}{\pi^{*(1)}(x)} \tag{224}$$

and its conditional expectation is

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1pm}\right] = \mathbb{E}_{2|1}\left[t_{\tilde{y}_1} + \sum_{x \in s_2} \frac{e_1^{(D)}(x)}{\pi^{*(1)}(x)}\right] \tag{225}$$

$$\mathbb{E}_{2|1}\left[\hat{t}_{y_1,1pm}\right] = t_{\tilde{y}_1} + \sum_{x \in s_1} \frac{e_1^{(D)}(x)}{\pi^{*(1)}(x)}, \tag{226}$$

where the first term, the total of model predictions in D and it does not contribute to covariance over first-phase samples, because it is a constant.

The conditional expectation of the two-phase estimator $\hat{t}^{(2)}_{y_2,2p}$ is

$$\mathbb{E}_{2|1}\left[\sum_{x \in s_2^{(2)}} \frac{y_2^{(D)}(x)}{\pi^{*(2)}(x)}\right] = \sum_{x \in s_1^{(2)}} \frac{y_2^{(D)}(x)}{\pi_1(x)}, \tag{227}$$

as derived in annex C.1, formula (186).

The conditional covariance is

$$\mathbb{C}_{1|2}\left[\hat{t}_{y_1,1pm}, \hat{t}_{y_2,2p}\right] = \mathbb{C}_{1|2}\left[\sum_{x \in s_2^{(1)}} \frac{e_1^{(D)}(x)}{\pi^{*(1)}(x)}, \sum_{x \in s_2^{(2)}} \frac{\dot{e}_2^{(D)}(x)}{\pi^{*(1)}(x)}\right], \tag{228}$$

where the first term stems from the fact that given the working model, the only source of variance are residuals. For the second term refer to (199).