

Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

Reverse-engineering a Hypothetical Raking Process for the Estimation of Mean Squared Error of Raked Small Area Estimates

by François Verret and Braedan Walker

Release date: September 8, 2025



Statistics
Canada

Statistique
Canada

Canada

Reverse-engineering a Hypothetical Raking Process for the Estimation of Mean Squared Error of Raked Small Area Estimates

François Verret and Braedan Walker¹

Abstract

Small area estimation is frequently used to produce estimates at a disaggregated level where direct survey estimation does not have sufficient sample to produce precise estimates. Often this is done using the area-level Fay-Herriot model, by assuming the direct estimates are independent under the design and have a known variance, and applying a smoothing process to the variance estimates of the direct estimates to better meet that last assumption. It is not rare that small area estimates are benchmarked/raked to aggregated level direct estimates. This article shows that wrongly assuming independence can have a big impact on the MSE of the raked estimates. Values of the covariances between direct estimates are thus required for good point and MSE estimates. Getting good estimates of those covariances is difficult given the small sample sizes in some areas. An original way of deriving values for those covariances, by reverse-engineering a hypothetical raking process, is presented.

Key Words: Benchmarking; Fay-Herriot model; Working covariance matrix.

1. Introduction

In recent years, Statistics Canada has made significant efforts to produce estimates at a more disaggregated level. Its Disaggregated Data Action Plan addresses how different groups have different life experiences by breaking down the data into sub-categories according to indigenous peoples, gender, racialized populations and persons with disability. Data is also broken down to the lowest possible level of geography. Statistics Canada has thus been increasing its use of Small Area Estimation (SAE) methodologies to produce statistics, notably for manufacturing and labour statistics, as well as considering the use of such methodologies in other fields. Area level modelling with the Fay-Herriot model (Fay & Herriot, 1979) is the most often used method since it integrates efficiently and easily both model and design random mechanisms. When total estimates are considered, survey practitioners may want the small area estimates to match some (potentially published) direct estimates at a more aggregate level. To that end, a reconciliation process such as raking (Dagum & Cholette, 2006) can be applied.

This is the case of monthly SAE of total employment in the Labour Force Survey (LFS), the focus of this paper, where estimates are obtained for a geographical partitioning of each of the ten provinces defined by Census Metropolitan Areas (CMA), Census Agglomerations (CA) and Self-contained Labour Areas (SLA). CMAs correspond to the largest cities of the country: those with a population size of 100,000 people or more and a core of 50,000 people or more. CAs correspond to smaller cities, those that are not part of a CMA and that have a population of at least 10,000 people. SLAs are functional areas composed of Census Subdivisions outside of CMAs and CAs grouped according to commuting patterns (OECD, 2020). Although the monthly LFS sample is large (more than 53,000 households sampled; Statistics Canada, 2020) and the survey produces good estimates for most CMAs, it is not designed to produce precise estimates for all the areas each month. Therefore, many small areas will have little to no sample. For these areas, the estimate might be 0 and should have poor quality. The corresponding variance estimate will also be of poor quality for the same reason. It is for this reason, and because excellent auxiliary information is available at the area level, that we rely on small area estimation methods.

¹François Verret, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (francois.verret@statcan.gc.ca); Braedan Walker, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (braedan.walker@statcan.gc.ca)

The Fay-Herriot (FH) model is used monthly and cross-sectionally for SAE and a raking process is applied cross-sectionally to provincial employment totals. The basic FH model assumes design-based independence between the area-level direct estimates (or equivalently independence between the sampling errors of the model). However, independence does not hold in the LFS total employment context because of weight calibration. An extension to the FH model where sampling errors are dependent thus needs to be fit. In order to fit the basic FH model, variances of the direct small area estimate need to be known. A common practice to improve upon these estimates is to apply a variance smoothing process involving a generalized variance function model. When extending the FH model to account for correlated sampling errors between areas, an additional difficulty arises when estimating the covariance terms on top of the variance terms given the small sample sizes in the domains. This paper describes a method where a hypothetical raking process (note: not the one applied to the FH SAE estimates) is reverse-engineered to get covariance terms such that the smoothed variances are preserved and that the elements of the resulting variance matrix when added agree with good quality variance estimates of direct estimates at an aggregate level.

The paper is structured as follows. Section 2 presents the extended FH SAE model (which includes the basic model when all covariance terms are zero) and the initial application of the basic model to estimate total employment in the LFS. Section 3 describes raking, its application to the LFS total employment small area estimates and demonstrates where the independence assumption shows signs of failure. Section 4 describes the reverse-engineering of the hypothetical raking process to obtain a full working design variance matrix and how point and MSE estimates are improved by using the working non-diagonal variance matrix. A conclusion is given in Section 5.

2. The (basic and) extended Fay-Herriot model

Under the extended FH model (with possibly dependent sampling errors), the finite population parameter vector of interest (of dimension M by 1, where M is the number of areas) is

$$\theta = X\beta + Bv,$$

where B is a diagonal matrix of known constants and $v \sim N(0, \sigma^2 I)$. In our application, we work with total estimates and $B = \text{diag}(N_1, \dots, N_M)$, where N_m is the number of people aged 15 and over in area m . The direct estimator is given by

$$\hat{\theta} = X\hat{\beta} + Bv + e,$$

where $e \sim N(0, \Psi)$ is independent of v . The composite estimator (or Empirically Best Linear Unbiased Predictor under the last model, EBLUP) is given by

$$\begin{aligned} \hat{\theta}^{\text{SAE}} &= X\hat{\beta} + \hat{\sigma}^2 B^2 \hat{\Sigma}_{zz}^{-1} (\hat{\theta} - X\hat{\beta}) \\ &= \hat{H}^T X\hat{\beta} + (I - \hat{H}^T) \hat{\theta}, \end{aligned}$$

where $\hat{\beta} = (X^T \hat{\Sigma}_{zz}^{-1} X)^{-1} X^T \hat{\Sigma}_{zz}^{-1} \hat{\theta}$, $\hat{\Sigma}_{zz} = \hat{\sigma}^2 B^2 + \Psi$ and $\hat{H}^T = I - \hat{\sigma}^2 B^2 \hat{\Sigma}_{zz}^{-1}$. For estimation purposes, the full matrix Ψ is assumed to be known. The basic FH model corresponds to Ψ diagonal, is the model that is the most often used in practice and is the one programmed in most software, including Statistics Canada's generalized estimation system G-Est (Estevao et al. 2023). The basic model was the first fitted for the LFS application.

For that application, the auxiliary information is made of counts of working age population (15 to 64 years of age) and of employment insurance beneficiaries (quadratic spline regression was done for the former). The design variance of the direct estimates, ψ_m , are assumed to be known, but are clearly not in practice. Estimates $\hat{\psi}_m$ are obtained with traditional design-based methodologies (using Rao-Wu (1988) bootstrap) and a smoothing process is applied to them to compensate for their lack of precision caused by small sample sizes by fitting a generalized variance function (Beaumont & Bocci, 2016). We thus get smoothed values $\tilde{\psi}_m$, which are assumed to be close enough to ψ_m . The variance of the model errors (σ^2) is estimated using Restricted maximum likelihood (REML) given the known design error variances. The EBLUP of area m under the basic model is given by

$$\hat{\theta}_m^{\text{SAE}} = \hat{\gamma}_m \hat{\theta}_m + (1 - \hat{\gamma}_m) \mathbf{x}_m^T \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_m = b_m^2 \hat{\sigma}^2 / (b_m^2 \hat{\sigma}^2 + \tilde{\psi}_m)$ (b_m being the m^{th} element of the diagonal of B) and $\mathbf{x}_m^T \hat{\boldsymbol{\beta}}$ is the synthetic estimator for area m . The composite estimator will thus be closer to the direct estimator if the model variance is large compared to the design variance and vice-versa.

The MSE matrix of the composite estimator is given in the Appendix (as a special case of the MSE matrix of the raked composite estimator) and has components $G_1(\sigma^2)$, $G_2(\sigma^2)$ and $G_3(\sigma^2)$. $G_1(\sigma^2)$ is the leading term representing the MSE had β and σ^2 been known, $G_2(\sigma^2)$ is the variance due to the estimation of β and $G_3(\sigma^2)$ is the variance due to the estimation of σ^2 . Whether the sampling errors are independent or not (i.e. whether Ψ is diagonal or not), when fitting the basic FH model we have that the m^{th} element of the diagonal of $G_1(\sigma^2)$ is $g_{1m}(\sigma^2) = \gamma_m^2 \psi_m + (1 - \gamma_m)^2 N_m^2 \sigma^2$. In any case, an estimate of the MSE is given by $G_1(\hat{\sigma}^2) + G_2(\hat{\sigma}^2) + 2G_3(\hat{\sigma}^2)$.

3. Raking of the Small Area Estimates

LFS direct provincial estimates of total employment are of good precision and disseminated. When aggregating the small area estimates (under the basic model) at the provincial level, we see in Table 3-1 that they can be a lot different from the provincial direct estimates in relative terms, especially for Prince Edward Island and Saskatchewan. The differences are of the same magnitude and systematic when looking at other months (not presented). This is an indication that the model is not perfect and a motivation to apply raking to the small area estimates.

Table 3-1
Provincial discrepancies between direct estimate of total employment and aggregate of small area estimates assuming independent sampling errors for the month of November 2021

Province	Direct Est.	SAE Est.	Difference	Relative difference (%)
NL	229,601	224,525	5,076	2.2
PEI	81,098	73,215	7,883	9.7
NS	477,326	483,855	-6,529	-1.4
NB	360,082	360,306	-224	-0.1
QC	4,380,216	4,361,616	18,600	0.4
ON	7,660,158	7,778,329	-118,171	-1.5
MB	669,344	652,554	16,790	2.5
SK	556,206	509,706	46,500	8.4
AB	2,298,973	2,277,357	21,616	0.9
BC	2,708,386	2,701,059	7,327	0.3
Canada	19,421,390	19,422,521	-1,131	0.0

Raking is designed to restore aggregation constraints in time series systems, for example, after seasonal adjustment. Although we are not in the time series context, we can use a cross-sectional raking framework to have the provincial sum of the SAE estimates match with the provincial direct estimate. When raking our small area estimates $\hat{\theta}_m^{\text{SAE}}$, we are looking for new estimates $\hat{\theta}_m^{\text{SAE, raked}}$ which minimise the following distance:

$$\sum_{m=1}^M \frac{(\hat{\theta}_m^{\text{SAE}} - \hat{\theta}_m^{\text{SAE, raked}})^2}{|c_m \hat{\theta}_m^{\text{SAE}}|},$$

under the constraint that the provincial sum of the new estimates corresponds to the direct LFS provincial estimates. The relative changes in individual estimates in raking can be controlled using the “alterability coefficients”, c_m . To limit the changes to the most precise small area estimates, we have chosen to define c_m to be the estimated relative root mean squared error (RRMSE) of $\hat{\theta}_m^{\text{SAE}}$: $c_m = \sqrt{\varphi_m / \hat{\theta}_m^{\text{SAE}}}$, where $\varphi_m = \text{mse}(\hat{\theta}_m^{\text{SAE}})$.

To measure the benefits of raking to provincial estimates, we applied our procedures to the May 2016 LFS and compared our estimates with those of the 2016 Census long form using the Average Absolute Relative Difference (AARD) between the two sets of estimates. For CMAs and CAs, the AARD of the direct estimates is 23.9%. The (unraked) FH estimates have an AARD of 6.3%. The raked estimates have an AARD of 6.5%. Small area estimation thus has a very beneficial effect on CMA and CA estimates, while raking has a marginal negative effect. Among the

SLAs, the AARD of the direct estimates is 70.4%. The AARD of the FH estimate is 26.6%. That of the raked estimates is 22.2%. Both small area estimation and raking thus have a beneficial impact for those areas.

To estimate the MSE of the raked estimates, we first tried to apply a parametric bootstrap that assumes independent sampling errors ($\mathbf{e} \sim N(\mathbf{0}, \Psi)$, with Ψ diagonal). The following steps were repeated for each bootstrap replicate:

1. Generate census values: $\theta_{m^*} \sim_{\text{ind}} N(\mathbf{x}_m^T \hat{\boldsymbol{\beta}}, N_m^2 \hat{\sigma}^2)$
2. Generate direct estimates: $\hat{\theta}_{m^*} \sim_{\text{ind}} N(\theta_{m^*}, \hat{\psi}_m)$
3. Apply FH small area estimation.
4. Apply raking to the estimates resulting from Step 3 to the corresponding bootstrap provincial direct estimate (derived from aggregating the proper bootstrap direct estimate values $\hat{\theta}_{m^*}$).

The MSE of the provincial aggregates of the $\hat{\theta}_m^{\text{SAE, raked}}$ is theoretically equal to the variance of the provincial direct estimate because of raking. The corresponding parametric bootstrap variance estimates should thus be close to the variance estimates of the direct provincial estimates. Under the independence assumption, we see big differences as seen in Table 3-2, especially for the smallest provinces. Thus, contrarily to the pre-raking case, it is important not to assume Ψ to be diagonal when raking is done. Additionally, the danger is that the estimates of the RRMSE of raked small area estimates risk being inflated as well.

Table 3-2

Provincial discrepancies between the CV of the direct estimate of total employment and the RRMSE of the aggregate of small area estimates assuming independent sampling errors for the month of November 2021

Province	Direct CV	RRMSE of aggregates of raked estimates	Ratio
NL	1.5%	5.4%	3.6
PEI	1.5%	8.0%	5.3
NS	1.2%	3.9%	3.3
NB	1.3%	3.8%	2.9
QC	0.6%	1.4%	2.3
ON	0.5%	1.0%	2.0
MB	0.9%	3.7%	4.1
SK	1.1%	4.8%	4.4
AB	0.8%	2.1%	2.6
BC	0.8%	1.8%	2.3

With the chosen alterability coefficients (and assuming the rmse of the composite estimates are known rather than estimated), the first term of the MSE of the raked estimates can be shown to be:

$$\begin{aligned}
 g_{1m}^{\text{raked}} = & \gamma_m^2 \psi_m + N_m^2 (1 - \gamma_m)^2 \sigma^2 + \left(\frac{\sqrt{\varphi_m}}{\sum_{j \in \Omega_m} \sqrt{\varphi_j}} \right)^2 \sum_{j \in \Omega_m} N_j^2 (1 - \gamma_j) \psi_j \\
 & + 2N_m \gamma_m \frac{\sqrt{\varphi_m}}{\sum_j \sqrt{\varphi_j}} \sum_{j \neq m, j \in \Omega_m} N_j (1 - \gamma_j) \psi_{jm} + \left(\frac{\sqrt{\varphi_m}}{\sum_j \sqrt{\varphi_j}} \right)^2 \sum_{j \neq k; j, k \in \Omega_m} N_j N_k (1 - \gamma_j)(1 - \gamma_k) \psi_{jk},
 \end{aligned}$$

where Ω_m is the set of areas in the same province as m . The first two terms of the equation are those associated with the EBLUP. The third term is always positive and represents the added variance coming from raking had the sampling errors been independent. The last two terms are negative if the sampling errors are negatively correlated and partly explain the overestimation observed in our naïve/standard parametric bootstrap application.

The sampling covariance present in the direct estimates of the LFS is due to the estimation procedures. LFS strata are numerous and small, and this would guarantee Horvitz-Thompson (i.e. design weighted) estimators are independent. However, final calibrated LFS estimators are dependent because the calibration adjustment of the LFS weights is done

at the provincial level. The dependence is strong for total employment because it is relatively close to some calibration totals (i.e. the number of people aged 15 years and over by province). The FH model extended to dependent sampling errors should thus be fitted instead. The challenge is to get the covariance terms of Ψ on top of the variance terms given the small sample sizes.

4. Reverse-engineering of a hypothetical raking process to obtain the full sampling error variance matrix

Our first attempt was to assume that all area-level direct estimates within the same province are equally correlated. However, the resulting matrix is not necessarily a proper covariance matrix because it might not be positive semi-definite (it never was in our application). Doing a spectral decomposition of that matrix and equating negative eigenvalues to 0 gave semi-definite positiveness, but at the expense of keeping the diagonal/variance terms intact.

The desired properties of the $\tilde{\Psi}$ matrix are the following:

1. the diagonal should correspond to the smoothed direct variance estimates $\tilde{\psi}_m$;
2. the covariance structure should respect the covariance structure of some aggregate totals. In our case, we will respect the variance-covariance structure of the provincial estimates for the LFS since both LFS weight calibration and raking of the FH estimates are done at that level;
3. the matrix should be a proper covariance matrix. In particular, it should be symmetric positive semi-definite.

The key idea to construct such a matrix is that proper variance matrices appear naturally as a result of a statistical process. Inspired by the LFS sampling and provincial calibration of the LFS weights, we thus assume our direct estimates result from raking hypothetical independent variables at the provincial level and we use the variance matrix of those raked estimates as $\tilde{\Psi}$.

Let $X_m \sim \text{ind}(\mu_{X,m}, \sigma_{X,m}^2)$, $m = 1, \dots, M$ and $\underline{Y} \sim (\mu_Y, \Sigma_Y)$ be independent from one another, and be respectively hypothetical independent small area level normal variates and a vector of aggregate totals for which we have a good estimate of their covariance matrix (e.g. provincial estimates). We make the assumption that our direct estimates $\hat{\theta}$ are the result of raking \underline{X} to \underline{Y} . That is, in raking we obtained $\hat{\theta}$ by minimizing

$$\sum_{m=1}^M \frac{(X_m - \hat{\theta}_m)^2}{|c_m X_m|}$$

under the constraint $G\hat{\theta} = \underline{Y}$. The explicit solution to the raking problem (i.e. the vector of direct estimates) is given by

$$\hat{\theta} = \underline{X} + V_e G^T (G V_e G^T)^{-1} (\underline{Y} - G \underline{X}),$$

where $V_e = \text{diag}(|c_1 X_1|, \dots, |c_M X_M|)$.

One must choose the alterability coefficient c_m of the hypothetical raking. In general, in raking one is not restricted in this choice. However, in the present situation one will want to use a value that guarantees the reverse-engineering problem is solvable. In the usual raking context, one would most often define $c_m = \sqrt{V(X_m)}/X_m = CV(X_m)$, but to simplify equation solving that ensues, we defined $c_m = \sqrt{V(\hat{\theta}_m)}/X_m$. Thus $V_e = \text{diag}\left(\sqrt{\tilde{\psi}_1}, \dots, \sqrt{\tilde{\psi}_M}\right)$.

We also have

$$\begin{aligned} \tilde{\Psi} &= V(\hat{\theta}) = V\left(\underline{X} + V_e G^T (G V_e G^T)^{-1} (\underline{Y} - G \underline{X})\right) \\ &= V\left([I - V_e G^T (G V_e G^T)^{-1} G] \underline{X} + V_e G^T (G V_e G^T)^{-1} \underline{Y}\right) \\ &= V(A \underline{X} + C \underline{Y}) \\ &= \text{Adiag}(\sigma_{X,1}^2, \dots, \sigma_{X,M}^2) A^T + C \Sigma_Y C^T. \end{aligned}$$

The diagonal values of $V(\hat{\theta})$ are known, as well as A and Σ_Y . Using the diagonal elements of the last equation, one can thus solve to find the values of $\sigma_{X,1}^2, \dots, \sigma_{X,M}^2$. Let A^2 be the matrix of squared elements of A . Using linear regression theory, we can show that to minimize changes to the diagonal elements of $V(\hat{\theta})$ we should set:

$$\underline{\sigma}_X^2 = (A^{2T} A^2)^{-1} A^{2T} [D(V(\hat{\theta})) - D(C \Sigma_Y C^T)],$$

where D is the operator that takes the diagonal of a matrix and puts it into a vector. Those values can then be used in the previous equation to get the full $\tilde{\Phi}$ matrix. The matrix is by construction a proper covariance matrix. Furthermore, it has the desired diagonal (or minimizes changes to it), as well as the embedded Σ_Y covariance matrix for the targeted aggregates.

Table 4-1
Provincial discrepancies between direct estimate of total employment and aggregate of small area estimates not assuming independent sampling errors for the month of November 2021

Province	Direct Est.	SAE Est.	Difference	Relative difference (%)	Relative difference (%) when independence was assumed from Table 3-1
NL	229,601	231,945	-2,344	-1.0	2.2
PEI	81,098	78,809	2,289	2.8	9.7
NS	477,326	482,921	-5,595	-1.2	-1.4
NB	360,082	363,074	-2,992	-0.8	-0.1
QC	4,380,216	4,396,197	-15,981	-0.4	0.4
ON	7,660,158	7,722,312	-62,154	-0.8	-1.5
MB	669,344	667,764	1,580	0.2	2.5
SK	556,206	546,322	9,884	1.8	8.4
AB	2,298,973	2,295,764	3,209	0.1	0.9
BC	2,708,386	2,722,514	-14,128	-0.5	0.3
Canada	19,421,390	19,507,622	-86,232	-0.4	0.0

Using that covariance matrix in point estimation, the need for raking is not as striking as seen in Table 4-1, although the differences are still consistent from month to month (not presented). Using it in the MSE equations, we naturally get RRMSEs for the provincial aggregates of raked FH estimates that agree with the CVs of the direct provincial estimates as seen in Table 4-2. More importantly, the quality indicators of the raked small area estimates are not artificially inflated anymore, which is the main concern.

Table 4-2

RRMSE of the aggregate of small area estimates and average RRMSE of raked small area estimates under independence and non-independence assumptions for the month of November 2021

Province	CV of direct provincial estimate	RRMSE of aggregates of raked estimates		Average RRMSE of raked estimates	
		assuming Ψ diagonal	not assuming Ψ diagonal	assuming Ψ diagonal	not assuming Ψ diagonal
NL	1.5%	5.4%	1.5%	14.5%	9.4%
PEI	1.5%	8.0%	1.5%	16.0%	7.0%
NS	1.2%	3.9%	1.2%	8.8%	6.3%
NB	1.3%	3.8%	1.3%	10.2%	7.3%
QC	0.6%	1.4%	0.6%	8.9%	7.1%
ON	0.5%	1.0%	0.5%	7.0%	6.1%
MB	0.9%	3.7%	0.9%	12.7%	7.0%
SK	1.1%	4.8%	1.1%	12.7%	6.8%
AB	0.8%	2.1%	0.8%	7.8%	6.5%
BC	0.8%	1.8%	0.8%	9.2%	7.3%

5. Conclusion

When covariance of sampling errors is strong, one should use the FH model extended for non-independent sampling errors instead of the basic FH model. Using the proper covariance structure is key to obtaining the real EBLUPs and to having MSE components that have the behavior outlined in Prasad & Rao (1990) (i.e. G_1 being the leading term of the MSE if the number of areas is large enough). One key input is the variance matrix of the design errors Ψ . We have presented how to reverse-engineer a hypothetical raking process to get a proper working variance matrix that has the desired variances at the small area level and some aggregate levels.

Although the MSE estimates are improved compared to those obtained under the assumption of independence, their quality depends on how well we approximate the covariance terms in Ψ . This can be improved by adding constraints in the hypothetical raking (e.g. for the small areas with a large enough sample size or sampling fraction). The effect of the discrepancies could be studied in a simulation context.

Acknowledgements

We would like to thank Jean-François Beaumont, Keven Bosa and Cynthia Bocci for their support and contribution.

References

- Beaumont, J.-F., and Bocci, C. (2016), "Small Area Estimation in the Labour Force Survey", Paper presented at Statistics Canada's Advisory Committee on Statistical Methods, March 31, 2016.
- Dagum, E.B., and P. Cholette (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, New York: Springer-Verlag, Lecture Notes in Statistics, 186.

- Estevao, V., You, Y., Hidioglou, M., Beaumont, J.-F. and Rubin-Bleuer, S. (2023), “Small Area Estimation-Area Level Model with EBLUP Estimation- Methodology Specifications”, Statistics Canada document.
- Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2000), “Estimation of Census Adjustment Factors”, *Survey Methodology*, 26, pp. 31-42.
- OECD (2020), Delineating Functional Areas in All Territories, OECD Territorial Reviews, OECD Publishing, Paris, <https://doi.org/10.1787/07970966-en>.
- Prasad, N.G.N. and Rao, J.N.K. (1990), “The Estimation of the Mean Squared Error of Small Area Estimators.” *Journal of the American Statistical Association*, 85, pp. 163-171.
- Rao, J.N.K., and Molina, I. (2015), *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Rao, J.N.K., and Wu, C.F.J. (1988), “Resampling inference with complex survey data”, *Journal of the American Statistical Association*, 83, pp. 231-241.
- Statistics Canada (2020). Guide to the Labour Force Survey, Catalogue no. 71-543-G. <https://www150.statcan.gc.ca/n1/pub/71-543-g/71-543-g2020001-eng.htm>.

Appendix

The MSE of the raked composite estimator can be derived by following the steps of Isaki, Tsay & Fuller (2000):

1. Express $\hat{\theta}^{\text{raked}} - \theta$ as a function of $\hat{\beta}$, $\hat{\sigma}^2$, v and e
2. Linearize the result in terms of $\hat{\beta}$ and $\hat{\sigma}^2$ and express it as a function of e and $z = Bv + e$
3. Assume $\hat{\sigma}^2$ and z are independent and derive the approximate variance of $\hat{\theta}^{\text{raked}} - \theta$

This will be done in the following, assuming matrix V_e related to raking alterability coefficients is not random. The MSE of the (unraked) composite estimator corresponds to $A = I$ in the derivations, where $A = I - V_e G^T (G V_e G^T)^{-1} G$.

Step 1: Express $\hat{\theta}^{\text{raked}} - \theta$ as a function of $\hat{\beta}$, $\hat{\sigma}^2$, v and e

$$\begin{aligned}\hat{\theta}^{\text{raked}} &= \hat{\theta}^{\text{SAE}} + V_e G^T (G V_e G^T)^{-1} G [\hat{\theta} - \hat{\theta}^{\text{SAE}}] \\ &= A \hat{\theta}^{\text{SAE}} + (I - A) \hat{\theta} \\ &= A \hat{H}^T X \hat{\beta} + (I - A \hat{H}^T) (X \beta + Bv + e).\end{aligned}$$

Consequently

$$\begin{aligned}\hat{\theta}^{\text{raked}} - \theta &= A \hat{H}^T X \hat{\beta} + (I - A \hat{H}^T) (X \beta + Bv + e) - (X \beta + Bv) \\ &= A \hat{H}^T [X(\hat{\beta} - \beta) - z] + e.\end{aligned}$$

Step 2: Linearize the result in terms of $\hat{\beta}$ and $\hat{\sigma}^2$ and express it as a function of e and $z = Bv + e$

The constant term of the Taylor linearization is $e - A H^T z$, where $H^T = I - \sigma^2 B^2 \Sigma_{zz}^{-1}$.

The term in $(\hat{\beta} - \beta)$ is $\left. \frac{\partial \hat{\theta}^{\text{raked}} - \theta}{\partial \hat{\beta}} \right|_{\hat{\beta}=\beta, \hat{\sigma}^2=\sigma^2} (\hat{\beta} - \beta) = A H^T X (\hat{\beta} - \beta)$.

The term in $(\hat{\sigma}^2 - \sigma^2)$ is

$$\begin{aligned}\left. \frac{\partial \hat{\theta}^{\text{raked}} - \theta}{\partial \hat{\sigma}^2} \right|_{\hat{\beta}=\beta, \hat{\sigma}^2=\sigma^2} (\hat{\sigma}^2 - \sigma^2) &= -A \left. \frac{\partial \hat{H}^T}{\partial \hat{\sigma}^2} \right|_{\hat{\sigma}^2=\sigma^2} z (\hat{\sigma}^2 - \sigma^2) \\ &= A H^T B^2 \Sigma_{zz}^{-1} z (\hat{\sigma}^2 - \sigma^2).\end{aligned}$$

We thus have $\hat{\theta}^{\text{raked}} - \theta \cong e - A H^T z + A H^T X (\hat{\beta} - \beta) + A H^T B^2 \Sigma_{zz}^{-1} z (\hat{\sigma}^2 - \sigma^2)$.

Step 3: Assume $\hat{\sigma}^2$ and z are independent and derive the approximate variance of $\hat{\theta}^{\text{raked}} - \theta$

$$\begin{aligned}G_1(\sigma^2) &= V(e - A H^T z) \\ &= \Psi - \Psi \Sigma_{zz}^{-1} \Psi + (I - A) \Psi \Sigma_{zz}^{-1} \Psi (I - A)^T \\ G_2(\sigma^2) &= A H^T X V(\hat{\beta}) X^T H A^T \\ &= A H^T X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T H A^T \\ G_3(\sigma^2) &= A H^T B^2 \Sigma_{zz}^{-1} B^2 H A^T V_{\sigma\sigma}.\end{aligned}$$

If $\hat{\sigma}^2$ is obtained using REML, then $V_{\sigma\sigma} = \frac{2}{\text{trace}(P^2)}$, where $P = \Sigma_{zz}^{-1} - \Sigma_{zz}^{-1} X (X^T \Sigma_{zz}^{-1} X)^{-1} X^T \Sigma_{zz}^{-1}$ (Rao & Molina, 2015).

Note that since $GA = 0$ we have as expected $V\left(G(\hat{\theta}^{\text{raked}} - \theta)\right) = G\Psi G^T + 0 + 0$, i.e. the design-variance matrix of the raking totals. Similar derivations when independence of the sampling error is wrongly assumed give

$$G_1(\sigma^2) = (I - AH_d^T)\Psi(I - AH_d^T)^T + \sigma^2 AH_d^T H_d^{\square\square} A^T,$$

where $H_d^T = I - \sigma^2 B^2 D_{zz}^{-1} = \text{diag}(1 - \gamma_1, \dots, 1 - \gamma_M)$ and D_{zz} is a diagonal matrix with the same diagonal elements as $\sigma^2 B^2 + \Psi$. In the absence of raking (i.e. when $A = I$), we thus have $g_{1m}(\sigma^2) = \gamma_m^2 \psi_m + (1 - \gamma_m)^2 N_m^2 \sigma^2$, as when the sampling error are rightly assumed independent.