Catalogue no. 75-005-M
ISSN 2292-3780
ISBN 978-0-660-73231-2

Labour Statistics: Technical Papers

Small Area Estimation Methodology Using Labour Force Survey Data

By François Verret, Braedan Walker, Cynthia Bocci and Jean-François Beaumont

Release date: September 17, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Small Area Estimation Methodology Using Labour Force Survey Data

by **François Verret, Braedan Walker, Cynthia Bocci** and **Jean-François Beaumont**

Statistics Canada's Labour Force Survey (LFS) produces estimates of labour force characteristics for different levels of geography across Canada. While standard weighted estimates, *i.e.*, direct estimates, can be produced from the LFS, these estimates are less precise for many small communities and rural areas due to the small sample size in those areas.

As part of the Disaggregated Data Action Plan (DDAP), a Small Area Estimation (SAE) methodology has been developed to produce monthly estimates of labour force characteristics for more detailed levels of geography than are available directly from the LFS, namely Census Metropolitan Areas (CMA), Census Agglomerations (CA), and a complementary geography called Self-contained Labour Areas (SLA). SLA are functional areas composed of Census Subdivisions (CSD) grouped according to commuting patterns (OECD, 2020).

1. Introduction

In the Labour Force Survey (LFS), direct estimates for a given area are obtained by using survey data from that area along with a set of suitable survey weights. Direct estimates are thus reliable if the sample size in the area of interest is large enough.

Small area estimation (SAE) methods attempt to produce reliable estimates when the sample size in the area is small. This is achieved by complementing the small amount of survey data with additional information that takes the form of a model involving auxiliary data. Small area estimates of employment counts and unemployment rates were produced using an "area-level" model. An area-level model is a set of assumptions about the relationship between direct estimates and auxiliary data at the area level. In this application, a linear model with normal errors was considered. The resulting small area estimate for a given area is a linear combination of the direct estimate for that area and a prediction from the model. The latter is often called a synthetic estimate and involves survey data from the area of interest as well as from other areas used in the modelling. As a result, small area estimates are indirect estimates in the sense that they use survey data outside the area of interest. For the areas with the smallest sample sizes, the direct estimates are not reliable and the small area estimates are driven mostly by the predictions from the model. However, for the areas with the largest sample sizes, it is the opposite and the small area estimates tend to be close to the direct estimates.

As mentioned above, the use of an area-level model requires auxiliary data available at the area level. There is no need to have this information at the unit (person) level. However, it must come from a source independent of the sampling selection mechanism of the survey. In the LFS, the number of employment insurance beneficiaries in each area of interest was used as well as demographic projections of the number of persons aged 15 to 64 years and 65 years and older. The former is obtained from an administrative source and the latter are obtained from the Centre for Demography at Statistics Canada. The production of small area estimates depends on the availability of these auxiliary data for the reference period, and as such, the small area estimates are less timely than direct estimates.

Section 2 of this document describes the area-level model used to produce employment counts and unemployment rates for the CMA, CA and SLA. In section 3, diagnostics used for model validation and evaluation of small area estimates are briefly discussed.

The small area estimates are obtained independently each month using the small area estimation module of the generalized software G-EST version 2.03 (Estevao et al., 2023a, 2023b).

2. Area-level model

Suppose that we are interested in estimating a certain population parameter for a given area i (and for a given month), which we denote by θ_i. For instance, θ could be the employment count or unemployment rate in area i. A direct estimator of θ_i is denoted by $\hat{\theta}_i$. The sampling error is defined as $e_i = \hat{\theta}_i - \theta_i$. The Fay-Herriot model is the most common area-level model. It has two components: the sampling model and the linking model.

The sampling model can be expressed as:

$$\hat{\theta}_i = \theta_i + e_i \ ,$$

where $E_p(e_i) = 0$ and $\mathrm{var}_p(e_i) = \psi_i$. The subscript p indicates that the expectation and variance are taken with respect to the sampling design (or the sample selection mechanism). The implicit assumption is that the direct estimator is unbiased under the sampling design. This assumption seems reasonable in the LFS given the types of parameters considered and estimation methods used. The quantity ψ_i represents the variance of $\hat{\theta}_i$ with respect to the sampling design and is typically unknown. A direct estimator of ψ_i is denoted by $\hat{\psi}_i$. In the LFS, $\hat{\psi}_i$ is obtained using a bootstrap method.

The second component of the Fay-Herriot model is the linking model. A common linking model is:

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i \ ,$$

where $E_m(v_i|\mathbf{z}_i) = 0$, $\mathrm{var}_m(v_i|\mathbf{z}_i) = \sigma_v^2$, $\mathbf{z}_i$ is a vector of auxiliary variables for area i, and $\boldsymbol{\beta}$ and σ_v^2 are unknown model parameters. The definition of $\mathbf{z}_i$ for the employment count and unemployment rate is provided at the end of this section. The subscript m indicates that the expectation and variance are taken with respect to the model.

In addition to the above model assumptions, the errors e_i and v_i, $i = 1,...,M$, are usually assumed to be normally distributed and mutually independent. The quantity M is the number of areas (CMA/CA/SLA) used for modelling.

By combining the sampling and linking model we obtain the Fay-Herriot model:

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + a_i \ ,$$

where $a_i = v_i + e_i$, $E_{mp}(a_i|\mathbf{z}_i) = 0$, $\mathrm{var}_{mp}(a_i|\mathbf{z}_i) = \sigma_v^2 + \tilde{\psi}_i$ and $\tilde{\psi}_i = E_m(\psi_i|\mathbf{z}_i)$ is a smooth variance. The subscript mp indicates that the expectation and variance are taken with respect to both the model and sampling design. Assuming $\tilde{\psi}_i$ is known, estimates of the model parameters $\boldsymbol{\beta}$ and σ_v^2 can be obtained using the restricted maximum likelihood method and are denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$.

The synthetic estimate of θ_i is simply the predicted value $\mathbf{z}_i'\hat{\boldsymbol{\beta}}$. It is called an indirect estimate because $\hat{\boldsymbol{\beta}}$ is obtained using the direct estimate from area i, $\hat{\theta}_i$, as well as direct estimates from areas other than i. Because it uses more data, the synthetic estimate is typically more stable than the direct estimate, particularly for areas with a small sample size. However, the synthetic estimate relies on the proper formulation of the Fay-Herriot model and could be significantly biased if the Fay-Herriot model is misspecified. A more robust and even more stable estimator that was implemented in the LFS is the composite estimator

$$\hat{\theta}_i^{SAE} = \hat{\gamma}_i\hat{\theta}_i + (1-\hat{\gamma}_i)\mathbf{z}_i'\hat{\boldsymbol{\beta}} \ ,$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ (\hat{\sigma}_v^2 + \tilde{\psi}_i)$. This small area estimate is a weighted average of the direct estimate and the synthetic estimate. When the direct estimate is stable, $\tilde{\psi}_i$ tends to be small and the small area estimate is close to the direct estimate. When the direct estimate is unreliable, $\tilde{\psi}_i$ tends to be large and the small area estimate is closer to the synthetic estimate.

In practice, the smooth variance $\tilde{\psi}_i$ is never known and must be replaced by a suitable estimate, $\hat{\tilde{\psi}}_i$. For areas with a large sample size, the direct estimate $\hat{\psi}_i$ of ψ_i obtained using standard methods, such as the bootstrap, is typically an efficient estimate of the smooth variance $\tilde{\psi}_i$. This is what was used in the LFS for the large areas. However, for areas with a small sample size, the direct estimate $\hat{\psi}_i$ is usually unreliable. To address this issue, the variance estimate $\hat{\psi}_i$ was modelled and its predicted value was used as the estimate of the smooth variance. This brought stability in the estimation of $\tilde{\psi}_i$ at the expense of introducing another model. Greater detail on estimation of the smooth variance is given in Beaumont and Bocci (2016).

Three inputs need to be provided to the SAE system for each area in order to obtain small area estimates $\hat{\theta}_i^{SAE}$, $i = 1, \ldots, M$:

i. direct estimates $\hat{\theta}_i$;

ii. smooth variance estimates $\hat{\tilde{\psi}}_i$;

iii. a vector of auxiliary variables $\mathbf{z}_i$.

We describe below how the vector $\mathbf{z}_i$ was defined for the estimation of the unemployment rate and employment count.

Let N_i^{emp} and N_i^{unemp} be the number of individuals employed and unemployed in area i, respectively. The unemployment rate in area i is defined as $\theta_i = \dfrac{N_i^{unemp}}{N_i^{unemp} + N_i^{emp}}$, whereas the employment count in area i is $\theta_i = N_i^{emp}$. Three auxiliary variables were available: N_i^{eib} , the number of employment insurance beneficiaries in area i, N_i^{15-64} , the number of individuals aged 15 to 64 in area i, and N_i^{65+} , the number of individuals aged 65 or older in area i. Let us define the beneficiary rate as $Q_i = \dfrac{N_i^{eib}}{N_i^{15+}}$ and the 15 to 64 rate as $R_i = \dfrac{N_i^{15-64}}{N_i^{15+}}$. For the estimation of the unemployment rate, we used $\mathbf{z}_i = (1, Q_i, R_i)'$. The first component of the vector represents the intercept. We also used terms to fit a linear spline in both Q_i and R_i . For the estimation of the employment count, modelling considerations led us to proceed differently. Instead of modelling N_i^{emp} directly, we modelled the proportion of employed individuals, $\dfrac{N_i^{emp}}{N_i^{15+}}$, using the intercept term, Q_i and R_i in $\mathbf{z}_i$ as in the unemployment rate model, but we also used terms to fit a quadratic spline in both Q_i and R_i . The small area estimate of $\dfrac{N_i^{emp}}{N_i^{15+}}$ is then multiplied by N_i^{15+} to obtain the small area estimate of the employment count N_i^{emp} .

Finally, a raking process was applied to the small area estimates of the employment count N_i^{emp} to ensure that when aggregated at the provincial level the total corresponds to the direct provincial estimate of number of people employed. Note that the areas do not aggregate to the province exactly since some areas overlap two provinces. In these cases, the combination of areas closest to the geographical definition of the province was used for raking.

3. Evaluation of small area estimates

The accuracy of small area estimates depends on the reliability of the Fay-Herriot model. It is therefore essential to make a careful assessment of the validity of the model before releasing estimates. For instance, it is important to verify that a linear relationship holds between $\hat{\theta}_i$ and $\mathbf{z}_i$, at least approximately. A simple way to verify the linearity assumption is to graph the standardized residuals:

$$\hat{a}_i = \frac{\hat{\theta}_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

against the predicted values $\mathbf{z}_i'\hat{\boldsymbol{\beta}}$. The linear assumption is reasonable when the graph does not reveal any particular trend. The standardized residuals are key statistics that can also be used to verify other model assumptions such as the normality of the model errors. Indeed, a test based on standardized residuals was developed to identify outlier areas, *i.e.*, areas that do not follow the same model as the other areas. Outlier areas would typically result in extreme residuals. Our model assessments in the LFS did not reveal any obvious model misspecification or outlier area. This was reassuring as the quality of small area estimates is highly dependent on the validity of the Fay-Herriot model.

The Mean Squared Error (MSE) is the usual concept used to evaluate the gains of efficiency resulting from the use of the small area estimate $\hat{\theta}_i^{SAE}$ over the direct estimate $\hat{\theta}_i$:

$$\mathrm{MSE}(\hat{\theta}_i^{SAE}) = E_{mp}(\hat{\theta}_i^{SAE} - \theta_i)^2 .$$

The MSE is unknown but can be estimated. When no raking is done to the estimates, one can use the formulas given in Rao and Molina (2015). For the final raked estimates of employment counts, the appropriate MSE formulas were derived and used (Verret and Walker, 2025). Gains of efficiency over the direct estimate are expected when the MSE estimate is smaller than the smooth variance estimate $\hat{\tilde{\psi}}_i$ or the direct variance estimate $\hat{\psi}_i$. In general, the small area estimates in the LFS were significantly more efficient than the direct estimates, especially for the areas with the smallest sample size.

We also compared direct and small area estimates for May 2016 with Census 2016 labour estimates, which are based on a much larger sample size. On average, the small area estimates were considerably closer to the Census estimates than the direct estimates (Beaumont and Bocci, 2016).

To ensure the robustness of the models over time, the small area estimation models were developed and evaluated using more than two years of monthly data prior to release.

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.

Estevao, V., You, Y., Hidiroglou, M., Beaumont, J.-F. (2023a). Small Area Estimation-Area Level Model with EBLUP Estimation- Description of Function Parameters and User Guide. Statistics Canada document.

Estevao, V., You, Y., Hidiroglou, M., Beaumont, J.-F. and Rubin-Bleuer, S. (2023b). Small Area Estimation-Area Level Model with EBLUP Estimation- Methodology Specifications. Statistics Canada document.

OECD (2020). Delineating Functional Areas in All Territories, OECD Territorial Reviews, OECD Publishing, Paris, https://doi.org/10.1787/07970966-en.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Verret, F. and Walker, B. (2025). Reverse-Engineering a Hypothetical Raking Process for the Estimation of Mean Squared Error of Raked Small Area Estimates. *Proceedings of the International Methodology Symposium*, Statistics Canada.