OVERVIEW OF SMALL AREA ESTIMATION METHODS USED FOR THE ESTIMATION OF MEAN LIQUID ASSETS

October 23, 2020
Statistics Canada

This report summarizes the modelling investigations that were conducted to obtain small area estimates of the mean liquid assets for each Census Division (CD) in Canada. Two data sets were used to obtain the final estimates: the 2016 Survey of Financial Security (SFS) and the 2016 Census of Population. The report is divided into three sections. The first section simply describes the proposal that was agreed on between Statistics Canada and the Community Data Program. The second section provides some details on the data used, the choice of auxiliary variables and the preliminary models considered. Standard regression techniques are used to obtain the preliminary models. Then, in the third section, we describe the small area models that were considered to obtain small area estimates. Special estimation techniques are required to properly account for the data structure and the complex sampling design used in the SFS. Some of these techniques are available in the Small Area Estimation component of the Generalized Estimation System (G-Est) developed at Statistics Canada. The system G-Est was used to obtain the final small area estimates.

1. PROPOSAL AGREED UPON BETWEEN STATISTICS CANADA AND THE COMMUNITY DATA PROGRAM IN JULY 2020

Using data from the 2016 SFS and data from the Census of Population of 2016, small area estimates of mean liquid assets at the CD level will be produced by Statistics Canada for the year 2016.

Survey estimates of mean liquid assets at the CD level will be drawn from SFS 2016. In a second step, these estimates of mean liquid assets at the CD level will be modeled as a function of socio-economic characteristics of families defined at the CD level. To ensure an accurate measurement of these socio-economic characteristics at the CD level, Census 2016 data will be used.

Two versions of mean liquid assets estimates will be considered. The first version will be based on all observations from SFS 2016. The second version will be based on a subset of SFS 2016 from which families with extremely high wealth/assets will be excluded.

Small area estimates will be constructed at the CD level, rather than the CSD level or Census tract level, to maximize the likelihood that the small area methodology used will be successful.

Any additional work to be performed after the aforementioned project is completed will be conducted on a cost-recovery basis.

2. DATA AND METHODS FOR PRELIMINARY MODELLING OF LIQUID ASSETS

This section explains how survey estimates of mean liquid assets at the Census Division (CD) level—the dependent variable—are modelled as a function of auxiliary variables (regressors).

Mean liquid assets of economic families and persons not in economic families are estimated at the CD level, using the Survey of Financial Security (SFS) of 2016.

Liquid assets are defined as follows:
Liquid assets = deposits in banks + financial investments + 0.9* RRSPS ,

where financial investments include: a) stocks, b) bonds, c) mutual funds, and d) tax-free savings accounts. A 10% withholding tax is assumed when withdrawing RRSPs, hence the use of the factor 0.9 pre-multiplying RRSPs.

The SFS 2016 uses the 2016 boundaries for CDs. The SFS 2016 has an overall sample size of roughly 12,000 economic families and persons not in economic families.

In principle, estimates of mean liquid assets can be computed in SFS 2016 for 258 CDs. However, 53 of these CDs have sample sizes lower than 10 when the top 1% of liquid assets are excluded. As a result, when focusing on CDs which have 10 observations or more, mean liquid assets can be computed for only 205 (i.e. 258 – 53) CDs.

The Census of Population of 2016 is used to compute the auxiliary variables—the regressors—at the CD level. The regressors derived from the Census of Population of 2016 can be computed for 293 CDs. However, since SFS 2016 allows meaningful estimates of mean liquid assets for only 205 CDs, the modelling exercise is undertaken for these 205 CDs.

Two different survey estimates of mean liquid assets are computed:

$\hat{y}_{1r}$ = SFS estimate of mean liquid assets for the CD r (only CDs with 10 observations or more are considered)
$\hat{y}_{2r}$ = SFS estimates of mean liquid assets for the CD r, excluding the top 1% of the distribution of liquid assets (again, only CDs with 10 observations or more are considered)

The models used are the following:

$$\ln(\hat{y}_{1r}) = \beta_0 + \beta_1 \ln(x_{1r}) + \beta_2 x_{2r} + \varepsilon_r \qquad (1)$$
$$\ln(\hat{y}_{2r}) = \beta_0 + \beta_1 \ln(x_{1r}) + \beta_2 x_{2r} + \varepsilon_r \quad , \qquad (2)$$

where

x_{1r} is the mean family size-adjusted income after tax in CD r,

x_{2r} is the percentage of major income earners in CD r who are immigrants,

ε_r is the model error, and

β_0, β_1 and β_2 are model parameters.

The rationale for equations (1) and (2) is simple:

a) CDs with higher income are expected to have accumulated more savings and thus, to have greater liquid assets, on average, than other CDs.
b) Immigrants are generally attracted to economically dynamic regions, which are expected to have higher wealth holdings and liquid assets.

For these reasons, the model parameters β_1 and β_2 are expected to be positive.

Models including the mean age of major income earners in CD r or the percentage of major income earners with a bachelor's degree or higher education in CD r generally yield a lower adjusted R squared. The same is true for models in levels, as opposed to those in logs.

Readers are reminded that:

a) $\ln(\hat{y}_{1r})$ and $\ln(\hat{y}_{2r})$ are computed from SFS 2016;

b) $\ln(x_{1r})$ and x_{2r} are computed from the Census of Population of 2016.

Since the goal of the exercise is to treat CDs as the unit of analysis, equations (1) and (2) are unweighted, i.e. give the same weight to each of the CDs in the sample.[1] Equations (1) and (2) are estimated using ordinary least squares (OLS) methods.

Estimation results are shown in Table 1. As expected, $\ln(x_{1r})$ (= ln_income_r in Table 1) and x_{2r} (= p_immig_r in Table 1) are positively correlated with $\ln(\hat{y}_{1r})$ (= ln_y1_r in Table 1) or $\ln(\hat{y}_{2r})$ (= ln_y2_r in Table 1). Since the estimate of β_1 equals roughly 1.6, a 1% increase in CD-level mean income after-tax predicts a 1.6% increase in CD-level mean liquid assets.

Table 1: Modelling CD-level mean liquid assets, 2016

Dependent variable is:	**ln_y1_r**	**ln_y2_r**
Auxiliary variables	**parameter estimates**	
ln_income_r	1.613 (0.286)	1.555 (0.267)
p_immig_r	0.013 (0.005)	0.008 (0.004)
Adjusted R squared	0.2485	0.2207
Number of CDs	205	205

Note: See text for details. Standard errors are in parentheses.
Source: Statistics Canada, Survey of Financial Security of 2016 and Census of Population of 2016.

[1] However, the CD-level dependent variable and regressors were computed using sampling weights from either SFS 2016 or Census 2016.

3. SMALL AREA ESTIMATION MODELS

The weighted estimate of mean liquid assets in CD r obtained from the SFS, $\hat{y}_{1r}$ (or $\hat{y}_{2r}$), may be unreliable when the CD sample size is small. In other words, there is a non-negligible chance that the estimate $\hat{y}_{1r}$ is far from the true population mean of liquid assets in CD r, denoted by y_{1r}. Small area estimation methods aim at producing more reliable estimates in this situation by complementing the SFS data with model assumptions that link the SFS estimates to auxiliary data from the Census.

The linear Fay-Herriot model is by far the most commonly used in practice to obtain small area estimates. It is often represented by its two components: the sampling and linking models provided below:

Sampling model: $$\hat{y}_{1r} = y_{1r} + e_r \qquad (3)$$

Linking model: $$y_{1r} = \beta_0 + \beta_1 x_{1r} + \beta_2 x_{2r} + v_r \qquad (4)$$

The term e_r is called the sampling error whereas the term v_r is called the linking model error. By combining these two components, we obtain the linear Fay-Herriot model:

Fay-Herriot model: $$\hat{y}_{1r} = \beta_0 + \beta_1 x_{1r} + \beta_2 x_{2r} + \left(v_r + e_r\right)$$

The Fay-Herriot model looks like a standard linear regression model but its error structure is different and it is why special estimation methods are necessary to estimate its unknown parameters. We denote the estimates of the model parameters β_0, β_1 and β_2 by $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

An important task that Statistics Canada performed during the production of the small area estimates is the validation of assumptions underlying the Fay-Herriot model. The main assumptions that were validated are those of linearity and normality of the model error, $v_r + e_r$. Model validation was conducted by carefully examining model diagnostics and graphs, and making modifications until a suitable model has been found. In particular, a handful of outliers were identified through this validation process. These outliers were then excluded for the estimation of model parameters.

The application of the Fay-Herriot model requires knowing the variance of the SFS estimates (or square standard error). While these variances can be estimated using the bootstrap weights produced by the SFS, these SFS variance estimates may be quite unstable when the sample size in CDs is small. This instability was reduced by modelling the SFS variance estimates using a log-linear smoothing model. The resulting variance estimates are called smoothed variance estimates, and those were used in the application of the Fay-Herriot model. Greater detail about the log-linear smoothing model and its validation can be found in Hidiroglou, Beaumont and Yung (2019).

The objective of small area estimation methods is to estimate the true population mean of liquid assets, y_{1r}, for each CD observed in the Census. For any given CD for which an SFS estimate is available, the small area estimate is a weighted average of the SFS estimate, $\hat{y}_{1r}$, and the model prediction, $\hat{\beta}_0 + \hat{\beta}_1 x_{1r} + \hat{\beta}_2 x_{2r}$, also called the *synthetic* estimate. This weighted average is called the *composite* estimate of y_{1r}. This small area composite estimate leans towards the SFS estimate when the latter is precise, typically when the SFS sample size in CD r is large. However, when the quality of the SFS estimate is poor in a given CD, the composite estimate will lean towards the synthetic estimate. When the SFS estimate is not available in a given CD (no SFS sample in that CD), the small area estimate is simply the synthetic estimate. The standard error of the small area estimates is also produced. It provides an indication of the quality of the small area estimates. Another quality indicator often used is the Coefficient of Variation (CV) of an estimate. It is defined as the standard error divided by the small area estimate. We observed that all but one CD, out of the 293 CDs present in the Census, had a CV less than 24%, and 285 had a CV less than 20%. This is a significant improvement over the quality of the initial SFS estimates, which often had a larger CV than the CV of small area estimates.

The above Fay-Herriot model and corresponding smoothing model were also applied with the SFS estimates $\hat{y}_{2r}$. Gains in precision were slightly smaller than those obtained using $\hat{y}_{1r}$. For instance, all but one CD of the 293 had a CV less than 24%, and 276 had a CV less than 20%. Although these small area estimates were also produced along with their standard error, we do not recommend using them. We recommend using the small area estimates obtained with the SFS estimates $\hat{y}_{1r}$ as they are slightly more precise.

In Section 2, a logarithmic model was considered, whereas the Fay-Herriot model in this section is linear. It is possible to extend the Fay-Herriot model to a logarithmic version. In that case, the sampling model (3) is still appropriate but the linking model (4) is replaced with the following logarithmic linking model:

Linking model: $$\ln(y_{1r}) = \beta_0 + \beta_1 \ln(x_{1r}) + \beta_2 x_{2r} + v_r \qquad (5)$$

A more complex Hierarchical Bayes estimation methodology can be used to estimate the model parameters and obtain the required small area estimates. This logarithmic linking model was tested but did not yield superior gains in precision than those obtained with the linear Fay-Herriot model. Therefore, we do not recommend the logarithmic linking model, and these small area estimates are not included in the file provided to the Community Data Program.

The reader is referred to Rao and Molina (2015) or Hidiroglou, Beaumont and Yung (2019) for greater detail about the Small Area Estimation methods used in this project.

REFERENCES

Hidiroglou, M.A., Beaumont, J.-F., and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, **45**, 1, 101-126. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf.

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

APPENDIX 1: VARIABLE DEFINITIONS:

All variables are mean values or percentages computed at the CD level. All amounts are in 2016 dollars.

liquid1: mean liquid assets
liquid2: mean liquid assets, excluding the top 1% of the distribution of liquid assets

msaincat: mean (family-size) adjusted income after-tax of families and persons not in economic families

mage: mean age of major income earners
magesq: mean value of: age squared / 100

mhsless: percentage of major income earners with a high school education or less
mbaplus: percentage of major income earners with a bachelor's degree or higher education

mimmig = p_immig_r: percentage of major income earners who are immigrants
mrimmig: percentage of major income earners who are recent immigrants, i.e. who arrived in Canada in 2005 or afterwards.

melderly: percentage of major income earners who are aged 65 and over

APPENDIX 2: SAS CODES FOR THE REGRESSIONS IN TABLE 1:

```
rsubmit;
data out.explore_new; set out.analyze; where nobs2 >=10;

/* nobs2 is the number of observations per CD when the top 1% of the distribution of liquid assets is
excluded */

lny1 = log(liquid1);
lny2 = log(liquid2);
lnx  = log(msaincat);

proc reg data=out.explore_new; where nobs1>=10;
model lny1 = mimmig lnx;

proc reg data=out.explore_new; where nobs2>=10;
model lny2 = mimmig lnx;
run;
```