Catalogue no. 71-526-X
ISBN 978-0-660-24068-8

# Methodology of the Canadian Labour Force Survey

Release date: December 21, 2017

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** STATCAN.infostats-infostats.STATCAN@canada.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

**Depository Services Program**

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

## 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.

## Acknowledgement

Sincere thanks are due to the many people who contributed in various ways to this document.

Our gratitude goes first to the primary authors of the various chapters: Lihua An, Justin Francis, Guy Laflamme, Yves Lafortune, Scott Meyer, Elisabeth Neusy, Steven Thomas and Sylvia White. These people worked on different aspects of the 2015 redesign of the methodology of the Labour Force Survey, and thus were the best people to write about it.

Besides these, numerous people in Household Survey Methods Division, Labour Statistics Division, and elsewhere in Statistics Canada, were also involved in the planning, reviewing, verification, revision, translation and production of this document. Sincere thanks are due to all of them.

## Table of contents

# Methodology of the Canadian Labour Force Survey

## Chapter 1 Introduction and Overview

### 1.0 Introduction

This publication is a reference guide to the methodology of the Labour Force Survey (LFS). This guide will primarily focus on the methodology used for the ten provinces of Canada, though the LFS also covers the three territories. It describes all current survey steps and highlights the changes made during the most recent sample redesign.

A separate document called the *Guide to the Labour Force Survey* (Catalogue No. 71-543-G, available online) is a complement to this report and describes the concepts, definitions and data produced in the LFS.

### 1.1 Background

The LFS was created after the Second World War to meet an urgent need for reliable and timely data on the labour market that reflected the transition from a war-time economy to a peace-time economy. The survey was designed to produce estimates on employment and unemployment at the regional and national levels.

Conducted quarterly when it began in 1945, the LFS became a monthly survey in 1952. In 1960, the Interdepartmental Committee on Unemployment Statistics recommended that the LFS become the official tool for measuring unemployment in Canada. Once this recommendation was adopted, the demand for data increased, since users wanted a broader range of labour market statistics and, in particular, more detailed regional data. The range of estimates produced by this survey has grown considerably over the years, and today it provides a detailed portrait of the Canadian labour market.

### 1.2 LFS concepts and products

The LFS is the official source of monthly estimates of total employment and unemployment. The main monthly indicators published include the unemployment rate, the employment rate and the participation rate. The LFS is also one of the main sources of information on socio-demographic characteristics of the working-age population such as age, marital status, level of education and family status.

Employment estimates are produced at various levels such as by sex, age group, industry, occupation, educational attainment, and immigrant status. Statistics are also produced on characteristics such as length of job tenure, usual and actual hours worked, and employee wages. The questions asked by the survey make it possible to examine a wide variety of topical employment issues such as involuntary part-time employment, multiple job-holding, and absence from work.

Unemployment estimates are produced by industry, occupation, duration of unemployment, type of work sought, and activity before looking for work. Supplementary measures of unemployment are also produced annually to shed further light on the degree of labour market slack and the extent of hardship associated with joblessness. Information is also available on the recent labour market activity of persons currently not in the labour market. The *Guide to the Labour Force Survey* provides a complete description of the LFS questionnaire content.

In addition to providing national and provincial estimates, the LFS produces data for sub-provincial regions, such as Economic Regions (ERs), Employment Insurance Economic Regions (EIERs) and Census Metropolitan Areas (CMAs). The federal and provincial governments use LFS data to distribute funding and other resources to the different political and administrative jurisdictions.

The LFS standard estimates are published every month in *Labour Force Information* (Catalogue No. 71-001-X, available online). A wide variety of labour market data are also available through CANSIM, Statistics Canada's key socioeconomic database and electronic extraction system. There are more than 100 CANSIM tables representing several thousand chronological series that are updated either monthly or annually with new LFS data.

The LFS can produce much more information than what is regularly published. Specific tabulations can be produced on a cost-recovery basis. For more information about available survey products and services, please see Section 9 of the *Guide to the Labour Force Survey*.

## 1.3 General survey overview and document structure

In the provinces, the LFS is a monthly household survey providing a sample of individuals who are representative of the civilian, non-institutionalized population, 15 years of age or older. Excluded from the survey's coverage are: persons living on reserves and other Aboriginal settlements, residents of institutions, full-time members of the Canadian Forces and residents of regions that are extremely remote or of extremely low population density. These groups together represent an exclusion of approximately 2% of the population aged 15 and over.

These groups are excluded from the survey target population due to specific operational challenges or for conceptual reasons. For example, it would be difficult to interview members of the Canadian Forces who live in locations that are inaccessible to LFS interviewers (*e.g.,* aboard warships or in military camps and barracks). Residents of institutions (for example, inmates of penal institutions, patients in hospitals or nursing home residents) are excluded because the LFS is designed to measure the labour force participation in the *current* labour market and residents of institutions are for the most part not able to participate in the *current* labour market and are not economically active.

The survey uses a two-stage sample design. In the first stage, a sample of primary sampling units (PSUs) corresponding to geographical regions is selected. In each selected PSU, a sample of dwellings is drawn at the second stage. Households are identified within the selected dwellings and all individuals in the household who are part of the target population are selected for the survey. The dwellings selected remain in the sample for a period of six months. Outgoing dwellings are replaced by dwellings from the same PSU, or from a similar PSU if the previous PSU is retired and replaced. This sample design results in a five-sixths month-to-month sample overlap, which makes the design efficient for estimating month-to-month changes. The rotation of dwellings after six months prevents undue respondent burden for households that are selected for the survey. The high proportion of PSUs in common between samples twelve months apart makes the design efficient for estimating year-to-year changes. Chapters 2 and 3 provide more information on the sample design.

Data collection for the LFS is carried out in the week following the LFS reference week. Usually, the reference week contains the 15th day of the month. In 2015, about 88% of sampled households responded to the LFS questionnaire each month. The LFS interview is mandatory and takes an average of eight minutes. The data are collected using a computer-assisted interviewing system. Several collection methods are used, including in-person and telephone interviews, and an Internet questionnaire. More information on the collection strategy is presented in Chapter 4.

In the days following collection, the data are processed. Editing, imputation and weighting are performed, and quality indicators are derived. These steps are described in Chapters 5, 6, 7 and 8. Despite the large amount of data to process every month, Statistics Canada publishes the LFS estimates only ten days after the end of collection.

The LFS sample and frame are also used for many of Statistics Canada's other social surveys. This is described in Chapter 9. Several appendices covering special topics and survey reference materials are included at the end of this guide.

## 1.4 Changes introduced in 2015

Every ten years, after the decennial census of population, the LFS undergoes a sample redesign to account for the evolution of the population and labour market characteristics, to adjust to the current and expected needs of data users (in terms of statistical analyses), and to update the geographical information used to carry out the survey.

The most recent sample design was gradually introduced in January 2015 and was fully implemented by June 2015.

The 2015 redesign introduced a number of major changes to the methodology of the survey. These changes were introduced to reduce survey costs, use updated collection methods, and allow data users to compute and report design-based variance estimates on their own.

In this survey redesign, the primary sampling units were constructed from the Dissemination Areas defined for the 2011 Census. In addition to streamlining the work involved with the sample redesign, this change makes the LFS geography more standard, which helps in the comparison of estimates across surveys and in analysis involving

multi-level modeling. The sample allocation strategy was modified to use quality targets that prevent the allocation algorithms from automatically increasing sample sizes in areas of low unemployment. The changes made to the PSUs, allocation, and stratification are detailed in Chapter 2.

An innovation that was introduced with the 2005 design, the use of existing lists of addresses, has been expanded significantly in the 2015 design. Statistics Canada's residential address register (AR) has been incorporated into a new household survey frame service. The Dwelling Universe File (DUF) is an extraction of addresses from the AR which is now being used to produce the list of addresses for over 90% of the PSUs in the LFS sample. This reduces the work of field interviewers who would otherwise have to create the list of addresses by directly observing the neighbourhoods / PSUs in the LFS sample. The frame service also supplies telephone numbers that will help interviewers establish contact with sampled households. More information about the address register is given in Chapters 2 and 3.

The LFS has added a third collection method in 2015: eligible respondents can now complete the questionnaire using the Internet. This new strategy is discussed in Chapter 4.

The overall imputation strategy did not change, but the list of variables used to create the imputation groups for donor imputation was reviewed and updated to include industry. This change and other changes made to edit and imputation are discussed in Chapter 5.

Last but not least, a significant change was made in 2015 to variance estimation. Starting in January 2015, the bootstrap method has replaced jackknife as the variance estimation technique for the LFS. This allows users to compute and report design-based variance estimates for state-of-the-art analyses on their own. This important change is presented in Chapter 7.

## 1.5 Changes coming beyond 2015

A few additional changes are already on the agenda for the coming years.

In January 2016, the classification systems used to categorize industry and occupation on LFS data will be updated to more recent standards. Specifically, the currently used North American Industry Classification System 2007 (NAICS 2007) will be updated to the NAICS 2012 standard, and the currently used National Occupational Classification–Statistics 2006 (NOC-S 2006) will be updated to the NOC 2011 standard.

In the coming years, the systems used for collection, processing, estimation and tabulation of LFS data will be migrated to new corporate business processes to achieve cost savings while maintaining the highest standards of quality and timeliness.

# Chapter 2 Sample design

## 2.0 Introduction

As mentioned in Chapter 1, the objective of the Labour Force Survey (LFS) is to produce reliable and timely data on employment, unemployment and characteristics of the working-age population at various levels of geography. In theory, such data could be acquired through an administrative source, a census of the population, or a sample survey. As there is currentkly no administrative source available that can produce the required estimates, nor is it feasible to conduct a census and contact everyone of working age every month to determine their employment status, a sample of the population is contacted, and their responses are used to produce monthly labour force estimates.

The sample design consists of all the steps to be carried out when selecting a sample. It impacts the quality of estimates produced and the survey costs. Since a significant portion of a survey's budget is spent on data collection, the sample design tries to minimize collection costs while maximizing data quality.

This chapter describes the various strategies the LFS uses to achieve this objective in the ten provinces. First, Section 2.1 presents some basic concepts of survey theory that will be used throughout this chapter. Section 2.2 outlines the overall LFS sample design. The sample allocation is described in Section 2.3. Section 2.4 describes how the clusters are formed and Section 2.5 describes how they are stratified. Finally, Section 2.6 describes the sample selection process and rotation methodology.

## 2.1 Some basic survey theory concepts

This section presents some concepts required in order to understand the sample design that is described in the following sections. For further information, a conceptual overview of survey theory is available in Survey Methods and Practices (Statistics Canada (2003)). More technical details can be found in one of the many books on sampling theory (*e.g.*, Cochran 1977 or Särndal, Swensson and Wretman 1992).

Data collected from a sample survey are used to produce estimates for the *target population* - the group or population of interest. Selecting a sample requires a *survey frame*, which should correspond as closely as possible to the target population although practical constraints may prevent this. The LFS selects a *probabilistic sample*, *i.e.*, a subset of the population for which the surveyed units are selected at random. *Estimates* for the population are calculated based on the information provided from this sample.

Estimates can differ depending on which individuals are selected in the sample. Also, the estimate produced from a sample differs from the estimate produced if the entire population was interviewed. These types of differences are called *sampling errors*. Survey results also have other errors not associated with the sample design, called *non-sampling errors*.

Two important measures of sampling error are *bias* and *sampling variance*. Suppose that it is possible to select several different samples using the same sample design. For each sample, an estimate of the characteristic of interest (*e.g.*, the number of unemployed, average number of hours worked) can be produced from the observed data. Bias is the difference between the average of the estimates produced from all of the possible samples and the corresponding true value for the whole population. The variability between the sample estimates, or how different they are from one another, is the sampling variance.

Bias can be caused by a number of sources, such as an imperfect survey frame, the method used to produce the estimate, or survey nonresponse. The sample design can add bias when some regions are excluded from survey coverage (*e.g.* due to prohibitive collection costs). This error component can be difficult to measure in practice because the true value for the population is generally unknown.

Sampling variance measures the spread between the estimates produced from all the possible samples. It reflects the degree of precision of an estimate: the smaller the sampling variance, the more precise the estimate. Sampling variance can be estimated from a single observed sample, even though it reflects variability between many theoretical samples.

Other measures of variability are derived from sampling variance. *Standard error* is obtained by taking the square root of the sampling variance and is often used to determine a *confidence interval* or to carry out a *statistical test*. Standard error is an *absolute measure* of variation, since it is measured in the same units as the estimate. Another measure is the *Coefficient of Variation* (CV), which is defined as the standard error divided by the estimate. The CV is a *relative measure*, since it is unit-free and calculated relative to the estimate. A third measure is the *design effect*, a relative measure calculated by dividing the sampling variance of an estimate obtained under the survey design by the sampling variance of a Simple Random Sample (SRS) of the same sample size. It can be used to compare the effectiveness of one sample design to another. The smaller the standard error, confidence interval length, CV or design effect is, the more precise the estimate is.

The primary goal of an effective sample design is to reduce the sampling variance given limited budget and operational constraints. A more efficient sample design can obtain the same precision of the estimates (as measured by sampling variance) using a smaller sample size than another less efficient design. Similarly, given a fixed total sample size, a more efficient sample design has lower sampling variance than a less efficient sample design.

Several factors influence the sampling variance of an estimate. The most influential factors are the number of individuals in the population, the number of individuals in the sample, the sampling method used to draw the sample, the response rate, and the homogeneity of the characteristic of interest in the population. The size of the population cannot be controlled. Response rates can sometimes be changed by data collection, but usually not by the sample design. However, by controlling the number of individuals in the sample, the sampling method used to draw the sample, and the homogeneity within sampled groups, a more effective sample design can be obtained.

## 2.2 Overview of the sample design

The LFS uses a complex sample design. A more efficient sampling method would be simple random sampling (SRS), where units are selected at random from a list with equal probability. However, a simple random sample of individuals requires a list of all individuals in the target population, which may be difficult to obtain in practice. Also, operational constraints may prevent the feasibility of an SRS design requiring a more complex sampling method to be used.

For most LFS estimates the target population is all persons in Canada aged 15 and over. It is impossible to directly select a sample of such persons to interview since a complete and up-to-date list of persons residing in the ten provinces is not available. Instead of selecting persons directly, it is easier to select dwellings and then identify and interview persons living in the selected dwellings. Although a fairly complete and regularly updated list of dwelling addresses is now available (see Section 2.2.1), selecting dwellings by simple random sampling would lead to a sample that would be too geographically spread out. As a result, travel costs associated with in-person collection could be exceedingly high.

To reduce travel costs, the sample of dwellings is taken through two consecutive selection stages. This method is called two-stage sampling. In the first stage, the provinces are divided into geographic regions called clusters or primary sampling units (PSUs). A random selection of these PSUs makes up the first stage sample. In the second stage, for each selected PSU, a list of dwellings in the region is established either by an extraction from the Address Register (The Dwelling Universe File (Section 2.2.1)) or through field listing. A second-stage sample of dwellings is selected from these lists. Dwellings are the secondary sampling units (SSUs). All of the residents, who are part of the target population occupying the selected dwellings within the selected clusters, make up the LFS sample of persons. This two-stage selection method is more complex but reduces the geographic spread of the sampled persons by clustering them, thereby reducing costs.

Starting in January 2015, the sample design in Prince Edward Island (PEI) was changed to one-stage sampling. This means that dwellings are selected directly from a list, without any clusters. Section 2.5.6 has more information about the one-stage sample design in PEI.

In addition to the monthly estimates described in Section 1.2, the LFS produces change estimates between two given reference periods. To improve the quality of these estimates, it is preferable to increase the overlap between the samples of these two periods, which is only possible by keeping the same dwellings in the sample for several months. Unfortunately, when the sample overlap is increased, the burden imposed on respondents rises because they must participate in the survey several times. This increased burden could lead to a lower

response rate. On the other hand, in addition to improved quality, a bigger overlap also reduces survey collection costs, since it costs less to obtain a response in subsequent months than in the first month. Therefore, the sample overlap is a compromise between the quality of the change estimates and the cost of survey operations versus the burden imposed on respondents.

It was decided to keep each dwelling in the LFS sample for six consecutive months. Subject to this limitation, the maximum overlap of the sample between two consecutive months is five-sixths. Therefore, it is necessary to replace one-sixth of the sample of dwellings each month. To implement this strategy, the LFS PSU population is divided into six rotation groups[1], with a sample selected in each group representing the whole population. The first rotation group is initially contacted in January. These dwellings then remain in the sample until June inclusively. In July, all the dwellings in rotation group 1 are replaced by a new sample of dwellings from the same rotation group. The second rotation group is made up of the dwellings initially surveyed from February to July inclusively, and so on for the other rotation groups. The rotation pattern is illustrated in Figure 2.1 at the end of the chapter. More information about the rotation of the dwelling sample is provided in Section 2.6.4.

The strategy of overlapping rotation groups has some advantages. First, it allows for more effective processing and estimation methods (described in Chapters 5 and 6). It also permits a simple method for selecting a subset of the LFS sample for other Statistics Canada surveys. Since each rotation group represents the whole population, it is straightforward to build the sample for another survey by grouping together the dwellings from an appropriate number of rotation groups. Information on using the LFS survey frame and sample for other household surveys is given in Chapter 9.

### 2.2.1 The Dwelling Universe File and its impact on the sample design

For the sample design, it is important to know approximately how many dwellings and how many occupied dwellings (*i.e.*, dwellings that correspond to households of persons) are in the LFS population. The counts are used for PSU creation, sample allocation and stratification. For previous LFS designs, these counts came from the most recent Census. However, the most recent Census counts are from May 2011 while this redesign was phased in starting January 2015.

To have a more up-to-date count of the total number of dwellings in the population, the Dwelling Universe File (DUF) was used for this design. The DUF is an extraction from Statistics Canada's Address Register (AR) database that contains residential addresses (dwellings). It is updated quarterly, using the latest administrative files available and the results of field listing and verification.

For planning the redesign, the June 2013 extract of the DUF was used. Since the DUF did not identify occupied dwellings, the number needed to be estimated. This was done by multiplying the total number of dwellings on the DUF by the dwelling occupancy rate from the 2011 Census.

## 2.3 Sample allocation

As described in Chapter 1, the LFS is the official source of monthly estimates of total employment and unemployment. The LFS is also one of the main sources of information on socio-demographic characteristics of the working-age population such as age, marital status, level of education and family status.

The LFS produces data for a variety of geographic regions including National, Provincial and sub-provincial regions, such as Economic Regions (ERs), Census Metropolitan Areas (CMAs) and Employment Insurance Economic Regions (EIERs). The sample allocation step specifies the target number of households to select in each of these regions[2]. It is established to ensure that the sample can produce estimates that satisfy various LFS precision objectives. This is a crucial step because the subsequent steps depend on it, and it ensures that the survey resources are effectively used. More information on the Census related geographies used by the LFS can be found in the Census Dictionary (Statistics Canada (2012)).

1. They can also be called rotation panels. This is commonly referred to as a rotating panel survey design.
2. LFS samples dwellings and not households (occupied dwellings). However, the allocation is expressed in terms of a targeted number of households. This number is used to determine the sampling rate of households in a region. When that sampling rate is applied to a list of dwellings, the number of dwellings sampled should on average yield roughly the target number of households. The precision will depend on the precision of occupancy rate and coverage of the DUF.

As explained in Section 2.1, the number of units sampled has a direct impact on the quality of the estimates produced by the survey. Since the total sample size is fixed, too much sample assigned to a given region will produce estimates for that region that are of better quality than required by the survey objectives to the detriment of the data quality in other regions. The LFS produces estimates at various geographical levels (Canada, provinces, economic regions, *etc.*), so it is necessary to reach a suitable compromise for all these estimates when allocating the finite-budgeted sample.

In order to meet the survey objectives and maintain the overall efficacy of the survey design, the LFS sample is allocated in two steps. In the first step, sample funded by Statistics Canada is allocated. In the second step, additional sample funded by Employment and Social Development Canada (ESDC) is added. This two-step approach is based on the hypothesis that the Statistics Canada LFS budget is ensured over a long period of time, but that the funding from ESDC could fluctuate over time. Thus, each part of the sample should be allocated separately to meet the appropriate objectives for which it is funded. Table B.4 in Appendix B provides the LFS sample allocation based on various geographical units.

### 2.3.1 Allocation of the sample funded by Statistics Canada

The first step consists of allocating the sample funded by Statistics Canada (36,000 households) among the 10 provinces. Statistics Canada has established LFS quality objectives for the provinces, Economic Regions (ERs) and Census Metropolitan Areas (CMAs). All targets are based on estimates of the number of unemployed persons. This is because unemployment is an issue of high interest and because unemployment, being rarer than employment, takes more resources to measure at the same quality in terms of CV than employment would. The purpose of the allocation is to ensure that the sample will be able to meet these objectives.

To make adjustments to the sample allocation of the previous design, it was necessary to predict the precision of the estimates of the number of unemployed persons (monthly and three-month moving average) for each province, ER, and CMA for a given sample size. These predictions were based on a CV estimation model involving the sample size, the estimate of the number of unemployed persons, and the estimated variance of that estimate. This model is based on data from 80 previous months of LFS. Therefore, it is implicitly assumed that the new sample design will have comparable efficiency to the previous one. It is also assumed that response rates, vacancy rates, and the number of adults per household will remain constant over time. Model assumptions were validated by analyzing LFS trends over the last nine years. Using this model, it was possible to predict the impact of allocation changes on the quality of future unemployment estimates.

The allocation strategy for the sample funded by Statistics Canada is based on the following criteria:

- For each province, the CV of the monthly estimate of the number of unemployed should be less than 7%;[3]
- For each ER, the CV of the three-month moving average estimate of the number of unemployed should be less than 25%;
- The minimum sample size for each ER is 200;
- For CMAs that do not correspond well to EIERs, that is for five CMAs and Lethbridge[4], the CV of the three-month moving average estimate of the number of unemployed should be less than 25%;

In previous designs, it was expected that CVs should be below the target for all months. However, in practice this does not happen. Over a ten-year design, some months can be outliers for some domains, which is difficult to predict. Allocating a larger sample to control for outliers can be both difficult and potentially a waste of finite resources better spent in a different region. For the redesign, the sample was allocated such that the CV targets would be met for at least 90% of the monthly estimates over time.

Using the CV estimation model and non-linear programming, the 36,000 households were allocated to provinces, ERs and CMAs to meet the above constraints while minimizing the variance of national monthly estimate of the number of unemployed. Further details on allocation to the provinces, ERs, and CMAs follow.

3. There is a slight modification to the quality target when unemployment is low, explained further below.
4. It was expected that Lethbridge would become a CMA after the 2016 Census.

### Allocation to the ten provinces

For the provinces, the previous design aimed to have CVs less than 7% for the monthly estimates of the number of unemployed. In practice, this was not achieved for many provinces. Simulations showed that it would not be possible to meet a 7% CV target for all provinces and in all months using only the 36,000 sampled households funded by Statistics Canada. Given that the sample size was fixed, the only solution was to modify the targets.

As mentioned earlier, the CV is a relative quality measure. For very low levels of unemployment, CVs tend to be higher. On the other hand, standard error is an absolute quality measure, in that it is measured in the same units as the estimate. For a fixed sample size, CVs increase as unemployment rate decreases, even if the standard error remains the same. Because of a low unemployment rate in the Prairie provinces over the last few years, CVs were higher than the 7% target, even though standard errors were comparable to months with higher unemployment and lower CVs. A decision was made to adopt two allocation procedures: Allocate to provinces based on CVs when unemployment rate is above 5%; Allocate based on a comparable standard error size when unemployment rate is below 5%.

The exception to this rule was PEI. The sample size in PEI was kept at the same level as in the previous design. A sample increase was avoided because a further increase would have led to dwellings being selected a second time over the planned ten-year life of the new design, which would impose considerable burden on respondents.

### Allocation to Economic Regions

The CV target for ERs remained unchanged from the previous LFS designs. However, some changes were made to the geographic regions targeted since it is difficult to ensure that we meet these targets for all ERs. ERs that are small in terms of their household count were combined and the precision objective was applied to the combined ER. Four groups of ERs were combined with the last redesign. They were located in the northern regions of Quebec, Manitoba, Saskatchewan and British Columbia. For this redesign, three additional pairs of small rural ERs were combined in Newfoundland and Labrador, Manitoba and Alberta.

### Allocation to Census Metropolitan Areas

Ensuring that the sample from Statistics Canada supports CV targets for all Census Metropolitan Areas (CMAs) has been considered. However, simulations have shown that imposing such targets detract from the precision of more important provincial and national estimates. In most cases, the Census Metropolitan Areas correspond to Employment Insurance Economic Regions (EIERs). The additional ESDC sample allocated for EIER estimates (see Section 2.3.2) provides enough sample to have sufficient quality for most CMA estimates. Two CMAs (Moncton and Saint John) together form a single EIER and each has sufficient quality due to the ESDC sample.

There are six areas that do not correspond well to EIERs and where the CV requirements are not met even with the additional ESDC sample. Five of these areas are the CMAs of Peterborough, Barrie, Brantford, Guelph and Kelowna. The sixth area is Lethbridge. After each census, Statistics Canada reviews the list of CMAs. It was expected that Lethbridge would become a CMA following the 2016 Census, so it was important to ensure that the sample drawn in Lethbridge would be sufficient to produce good estimates after 2016. Additional sample was allocated to these six areas so they would have sufficient sample for the CVs of the three-month moving average estimate of the number of unemployed to be under 25%.

## 2.3.2 Allocation of the sample funded by ESDC

This step of the LFS sample allocation involves adding sample funded by ESDC to the core sample funded by Statistics Canada. LFS unemployment rate estimates for EIERs are used by ESDC to establish employment insurance eligibility criteria and duration of benefits in each region. To help improve precision of unemployment rate estimates for EIERs, ESDC pays for an additional sample of 16,600 households.

ESDC eligibility criteria and benefits are determined based on ranges of the unemployment rate. ESDC needs comparable precision of estimates in all EIERs, except when region's unemployment rates are in the highest and lowest ranges. The lowest range is 6% and lower; therefore, in regions with low unemployment, ESDC just needs to be able to determine that the unemployment rate is below 6%. Similarly, the highest range is

above 13%; therefore, in regions with very high unemployment, ESDC just needs to be able to determine that the unemployment rate is above 13%.

Before allocating the sample funded by ESDC, the number of units allocated to each EIER by the sample funded by Statistics Canada had to be determined. First, the sample in an ER or CMA was proportionally allocated to the ER-EIER-CMA intersections based on the size of each intersection. By summing the intersections, the Statistics Canada sample allocated to each EIER was determined. For the sample funded by ESDC, the target sample size in each EIER was based on the following criteria:

- For each EIER[5], the CV of the estimated unemployment rate by three-month moving average must be less than 15%. However, in regions with low unemployment, ESDC just needs sufficient quality to conclude that unemployment rate is below 6%. Therefore, for rates less than 4.8%, the sample size required is such that the standard error must be small enough to conclude the unemployment rate is less than 6%. The value of 4.8% was chosen because it is the maximum value for which the upper bound of the confidence interval on the unemployment rate will be 6% when the CV is 15%.
- The minimum sample size for each EIER is 500;
- The quality of the estimates produced for each EIER must be similar from one EIER to another.

Once again, non-linear programming was used to solve this problem. After allocating the sample funded by ESDC, the total sample size (Statistics Canada and ESDC) assigned to each EIER was allocated to the ER-EIER-CMA intersections, proportionally to the size of each intersection. This new allocation was then compared to the one used before the redesign to identify potential errors in the model used and to predict the effectiveness of the new design. Changes in regional sample sizes and predicted CVs were used to evaluate the new design.

After this final step of sample allocation, two parameters were produced: the inverse sampling ratio (ISR) and the number of sampled households required for each intersection. The inverse sampling ratio is the number of households in the intersection divided by the number of households allocated to the sample for the intersection. It is used to determine the size and number of design strata (see Section 2.5) and during the sample selection process (see Section 2.6).

## 2.4 Creation of PSUs

In Section 2.2, the basic design was presented where the LFS has a selection of clusters as primary sampling units (PSUs) followed by a selection of dwellings as the secondary sampling units (SSUs)[6]. The first step of the two-stage sample design is determining the geographic boundaries of the PSUs used for the first stage of sample selection based on size, shape and other factors.

The determination of the size and shape of the PSUs is a compromise between collection costs and the sample design's efficiency. From a cost perspective, collection is cheaper if the shape of PSUs is geographically compact and contiguous, reducing travel time between the sampled dwellings within the PSU. If PSUs are made too large, it is too costly for an interviewer to visit all the sampled dwellings frequently enough to get responses. On the other hand, small PSUs remain in the sample for less time, which increases costs associated with PSU replacement. Also, when the PSUs are small, many PSUs will need to be selected. The selected PSUs will be generally further away from each other, again increasing the travel costs.

From a design perspective, the LFS could select either a few PSUs with many dwellings selected in each or many PSUs with a few dwellings selected in each. The latter case leads to a more efficient survey design[7]. However, moving towards such a design negates the advantages of a clustered design.

To determine the ideal size of a PSU given the above considerations, two elements are necessary. The first is a tool to evaluate the sampling variance resulting from different scenarios. This tool can be built using census data. The second is a relatively accurate model to estimate the collection costs for different scenarios of PSU size and the number of dwellings selected per PSU. To build this model, detailed information on costs is needed. Due to the

5. The Northern Manitoba EIER is an exception. Despite having some CVs above 15%, this region has historically had unemployment rate far above 13%, so better quality was not needed to determine EI benefits. Thus, to avoid high collection costs in Northern Manitoba, the sample size was kept at the same level as in the previous design.
6. PEI, which has a one-stage design, does not have clusters
7. Taking this argument to the extreme, the ideal solution would be to create PSUs containing one dwelling each. This is equivalent to one-stage sampling. However, as discussed in Section 2.2, one-stage sampling is currently too expensive to implement outside of PEI

complexity and recent major changes in the collection strategy (see Chapter 4), it was virtually impossible to build a valid updated cost model. Therefore, it was not possible to re-evaluate the ideal size of the PSUs. The ideal size of the PSUs that was used in the two previous redesigns (200 households) was maintained and used once again as a target for this design.

Once a target PSU size was established, PSUs were built from the standard geographical unit of Dissemination Areas (DAs) from Census 2011. This has several advantages. Using DAs as a basis for PSUs removes the need to create new geography definitions for LFS, as was done in previous designs. This streamlines the PSU creation process and reduces redesign costs. For analysis, using standard geographical units helps in comparing estimates across surveys, linking to auxiliary data from Census and other sources, and multi-level modeling. Also, since many household surveys typically sample for regions defined by standard geographical units, using DAs as PSUs simplifies use of the LFS frame by other household surveys which includes the ability to update as new DAs are defined as will be the case with the 2016 Census.

Unfortunately, some DAs are too large and others are too small for the ideal constraint of 200 dwellings per PSU. Simulation showed that having a lot of variability in PSU size would increase sampling variance under the LFS sample selection strategy (see Section 2.6). Some variability was inevitable in order to stay close to standard geographical units. An acceptable range of 100 to 600 households per PSU was determined. DAs below this range were joined with other contiguous DAs to form larger PSUs. DAs above this range were split into smaller contiguous and compact PSUs at the level of Census 2011 Dissemination Blocks or block faces.

An exception to this rule was made in Toronto. Toronto contained many DAs with more than 600 households, yet many were difficult to split into smaller PSUs due to the presence of high-rise apartment buildings with more than 600 units. It is inconvenient to split a single building into more than one PSU. First, unit occupancy changes frequently, making it difficult to accurately control for the number of households in each piece when dividing up the building. Also, once an interviewer has established regular access to an apartment building, it is quite efficient to continue to collect from other units in the building. This means that this large PSU would not have the same high costs as a large PSU of detached houses. Therefore, in Toronto many PSUs were created containing between 600 and 1000 households. To avoid a negative impact on design efficiency, these PSUs were grouped together into special strata (see Section 2.5.4).

Once all PSUs were created, a detailed analysis was done to identify those that were far from an urban centre and would probably have a very high collection cost. Depending on the situation, these PSUs were either stratified separately (see Section 2.5) or excluded from the survey frame. Under the previous LFS design, less than 1% of households in Canada were excluded. For this redesign, approximately 100,000 additional households were excluded in northern areas of the ten provinces, bringing the rate of exclusions up to 1.5%. See Appendix B.1 for more details on excluded areas. Excluding persons belonging to the target population of a survey inherently introduces bias into the survey estimates; however, the cost required to cover these regions was deemed too high relative to the potential impact on estimates.

## 2.5 Stratification

Stratification is the process whereby the population is divided into homogeneous, mutually exclusive groups called strata, in order to improve the efficiency of the sample design. In many surveys, strata are defined based on geographic domains of interest. In the case of the LFS, strata are formed within each domain of interest: ER-EIER-CMA intersections. This extra stratification ensures that the survey can accommodate the rotation, allocation and selection constraints that are described in this chapter.

The first step is to determine how many strata are needed within a domain. Once the number of strata is determined, the strata can be defined based on geographic, socio-economic and efficiency constraints and the PSUs can be grouped into these strata. Stratification will improve the sample design's efficiency if the PSUs grouped together are homogeneous, meaning that the households therein have similar characteristics. Once this process is complete, a survey frame can be created, containing all of the PSUs and their corresponding strata.

### 2.5.1 Changes made during this redesign

Two past strategies, which were previously introduced to the LFS stratification methodology in order to reduce the costs associated with in-person collection and listing, were discontinued for this design. First, in isolated urban areas, a three-stage sample design had been used so that, in the first stage, only one of a group of population centres would need to be visited at a time. Second, the rural PSUs that were the most expensive for in-person collection (due to high vacancy, distance from urban centres or lack of road access) had been stratified separately and given a reduced sampling rate to minimize the number of in-person visits needed. These innovations reduced collection and listing costs, but also decreased the efficiency of the sample design. In addition, they could lead to shifts in some local industry employment figures, especially when a particular PSU with many workers associated with a given industry was replaced by another PSU in a different area with a different dominant industry. Now, with more interviews handled by telephone instead of in-person (see Chapter 4), the expanded exclusion of high-cost remote areas (mentioned in Section 2.4), and the reliance on the DUF to provide dwelling lists, the potential cost savings offered by these two innovations appeared less favourable compared to the loss of design efficiency and the impact on estimates; therefore, they were discontinued for this redesign.

For the redesign, two-stage sampling was used in all provinces except PEI. Other changes in the stratification methodology include: PEI being stratified differently to facilitate the new one-stage design (see Section 2.5.6), a new type of special stratum being introduced to deal with the large PSUs in Toronto (see Section 2.5.4), and the specific needs of the Canadian Community Health Survey (CCHS) being taken into account (see Section 2.5.3).

### 2.5.2 Stratum size

The size and number of strata in each ER-EIER-CMA intersection is determined based on the sample allocation, the number of PSUs to select in each stratum and the number of households to select in each PSU (also called the sample take or density factor). The allocation to each intersection was explained earlier. The number of PSUs is based on the rotation strategy where one sixth of the sample rotates every month. To implement this approach, it is preferable to select six PSUs (or sometimes twelve) in each stratum. Finally, past studies have determined that in order to improve the sample design's efficiency, the sample take for a PSU should be ten in rural strata, eight in urban strata, and six in strata covering the Montréal, Toronto and Vancouver CMAs. Selecting more households per PSU in the rural strata reduces the travel costs per unit for in-person collection. At the other end of the spectrum, selecting six households per PSU in the largest CMAs helps to increase the number of PSUs required in the sample, improving the precision of the estimates. This reduction of the sample take also increases the number of strata needed. More and smaller strata should lead to an increase in the homogeneity of PSUs within, which should also improve the efficiency of the sample design.

By combining these constraints (allocation requirements, six PSUs selected per stratum and a fixed number of households selected in each PSU), the size requirement of each stratum within an intersection, in terms of households, can be calculated as:

$$M_h = ISR \times 6 \times m_h^* \qquad (2.1)$$

where

$M_h$ is the number of households to group together in each stratum of an intersection.

$ISR$ is the inverse sampling ratio as established during the first two steps of the sample allocation.

$m_h^*$ is the number of households to select per PSU. As explained in the previous paragraph, this number varies by the population density of the region (rural, urban, three largest CMAs).

The number of strata needed in each region can be determined by dividing the number of households in a region by this result and rounding the result to the most appropriate integer. Usually, the strata within an ER-EIER-CMA intersection that this process creates are approximately the same size.

### 2.5.3 Adjustments to geographic boundaries

Using the stratum size expression described above, it was not possible to create strata in some small ER-EIER-CMA intersections. Consequently, these small intersections were combined with a neighbouring intersection. Combining was done so that the combined group respected the boundaries of the CMA or EIER as much as possible. This approach implicitly gives more importance to the estimates by CMA and by EIER than by ER; therefore, the efficiency of the sample design decreased at the ER level, but was maintained for the EIERs and CMAs. In cases where 2011 Census boundaries for a CMA no longer matched the boundaries for the EIER representing that city[8], the resulting small intersections were treated as if the EIER boundaries matched the CMA boundaries. After combining the small pieces, there were 120 intersections covering the ten provinces in which stratification occurred.

Outside CMAs, it was favourable to create separate strata for urban and rural areas for three reasons: rural strata have more households than urban strata (see Equation 2.1 in Section 2.5.2); persons residing in rural areas have different characteristics from those residing in urban areas; and stratification that respects these areas allows for the implementation of more appropriate collection strategies. In some cases, an urban or rural area is too small to create a stratum using the size determined in Section 2.5.2. In such cases, it is necessary to combine the area with a neighbouring urban area or a rural area. Each case was evaluated separately.

The CCHS is a regular user of the area frame, and they rely on the LFS stratification methodology to identify the PSUs for their sample. For the LFS redesign, it is beneficial to make some adjustments to support their needs and minimize the impact on the LFS. The CCHS samples at a much higher rate in some rural regions than the LFS does and their geographic domains do not always correspond well with LFS regions. In the past, the CCHS would simply select more dwellings within the limited PSUs that were available causing much faster PSU rotation than planned (see Section 2.6.4) or select more PSUs than required by the LFS. For the redesign, the solution was to create additional CCHS specific strata that would ensure that there were enough PSUs selected in those regions to meet the CCHS requirements. This has less impact on LFS operations.

After the adjustments to the geographic boundaries, PSUs can be grouped into strata. Some PSUs were assigned to special strata (see Section 2.5.4), and the remaining PSUs in each intersection were stratified geographically and then optimally (see Section 2.5.5).

### 2.5.4 Special strata

Special strata can be divided into two categories: those created to improve efficiency, and those created to target specific populations. The first category is used to group remote PSUs as well as PSUs with a large number of dwellings in Toronto. The second category of special strata helps target sub-populations of interest for analysts who use LFS data.

#### Strata used to group inconvenient PSUs

Two special strata were created to group inconvenient PSUs: remote strata containing PSUs that are geographically isolated and difficult for in-person collection, and strata in Toronto containing PSUs with a large number of dwellings. By grouping these PSUs, the rate at which these PSUs are selected can be controlled.

A significant part of Canada is inhabited by a small portion of the population. Collection costs are high in regions with a small population, while the impact of these regions on the main LFS estimates is relatively low. Such PSUs were identified using data from Census 2011 on population density, distances to urban centres and accessibility by road. If there were enough of these PSUs in a province, they were grouped together into a remote stratum. By assigning these regions to specific strata, the number of these PSUs selected in a given sample can be better controlled, thereby better controlling the assignment of LFS resources.

As mentioned in Section 2.4, Toronto contains many PSUs with between 600 and 1000 dwellings, unlike other LFS PSUs in the provinces. If PSUs with 1000 dwellings were stratified with PSUs with 100 dwellings, there would be a considerable increase in sampling variance given the sample selection strategy used. However, the design stays efficient if PSU sizes are relatively homogeneous within a stratum. Since these PSUs could not be split any further,

8. EIER geography definitions date back to 2000. Due to urban sprawl, many CMAs that were once EIERs have now grown beyond the EIER boundaries.

they were instead grouped together as special strata. That way, the variability of PSU sizes within Toronto strata is reduced leading to a more efficient design.

Table B.2 in Appendix B presents the number of households in the first-category special strata.

### Strata to target certain sub-populations

Three types of sub-populations were targeted by special strata: households with high income, Aboriginal people, and recent immigrants. For simplification, the terms high-income strata, Aboriginal strata, and immigrant strata will be used from now on, although this is technically incorrect since these strata do not only contain high-income households, Aboriginal people, or immigrants. High income strata were created in most large CMAs. Aboriginal strata were created in British Columbia, Alberta and Saskatchewan. Immigrant strata were created in Manitoba only.

Because the LFS samples clusters of dwellings, instead of persons directly, it is difficult to target these rare sub-populations, especially when they do not all live in the same neighbourhood. Even in a neighbourhood with a higher prevalence of the sub-population, many households still will not have any members of the sub-population, so a sample of these dwellings may not yield many more members than a sample from a different region. Since there is no better tool available within the constraints of the LFS design to target sub-populations, special strata can help by at least ensuring that a PSU with higher prevalence is selected.

For special strata to effectively cover a target population, they need to have a higher prevalence of the target population and represent a large proportion of the overall target population. However, even if they produce good estimates for their target population, special strata cannot be justified if their introduction leads to a significant decline in the quality of the main LFS estimates. In order to find a viable compromise, two guidelines from a study done for the last redesign were used. The first guideline states that the strata must be created based on the prevalence of specific characteristics. For example, it would be futile to create an immigrant stratum in northern Manitoba, where the proportion of immigrants is very low. The second guideline states that no more than 8% of a domain can be used to create each type of special stratum. This limitation guarantees that the creation of these strata will not have a major adverse effect on the main LFS estimates, based on a study conducted using 1996 and 2001 Census data.

Using these two guidelines, the special strata were created in sequence. For each category, they were created by identifying the PSUs with the highest prevalence of the sub-population of interest. PSUs were then continually added to strata in decreasing order of prevalence until the 8% limit for the domain was reached. Using this approach, these strata are not contiguous and may be quite spread out geographically.

High-income strata were created first. PSUs in a given CMA were first classified in descending order based on the proportion of households with an income over $150,000 based on the 2012 T1 Family File (T1FF) generated from 2012 tax returns received by Canada Revenue Agency[9]. PSUs at the top of this list were assigned to a high-income stratum until the stratum's pre-determined size had been reached (see Section 2.5.2). If the limit of 8% was not attained, another high-income stratum was created for the same CMA. In this way, high income strata were created for most CMAs.

To create the Aboriginal strata, the basic strategy had to be slightly modified. The high-income strata respect the CMA boundaries, but a significant number of Aboriginal people live outside these boundaries, so special strata were created separately in CMAs and outside CMAs. Furthermore, some ER-EIER intersections outside CMAs were too small to form an Aboriginal stratum, although several PSUs in these intersections had a high proportion of Aboriginal households. To remedy this problem, the Aboriginal strata created outside CMAs respect the boundaries of just the EIERs, rather than those of the ER-EIER intersections. Finally, PSUs already assigned to a remote stratum could not also be assigned to an Aboriginal stratum. Among the remaining PSUs in the combined intersections, PSUs were put into the strata in descending order based on the proportion of households with at least one person who reported having an Aboriginal identity on the 2011 National Household Survey until the 8% limit was reached. As with high income strata, multiple strata were created where necessary to reach the 8% limit.

9. Since a T1FF record is not available for each household in a PSU, the proportion of high income households had to be estimated from the proportion among households with T1FF records in the PSU. In the majority of PSUs, T1FF records were available for 90% of households or more, so this estimate should be reliable. PSUs where only a few T1FF records were available were excluded.

To create the immigrant strata in Manitoba, strata also had to be created both inside and outside CMAs. Since most of the recent immigrant population of Manitoba resides in Winnipeg and the prevalence of immigrants was low elsewhere, only two immigrant strata were created outside Winnipeg. For these two strata, PSUs outside Winnipeg were put into descending order based on the proportion of households with at least one person who had immigrated to Canada in the last ten years according to the 2011 National Household Survey.

Using only 8% of Winnipeg and strata outside Winnipeg would have provided inadequate representation of the target population, since recent immigrants are disproportionately located in Winnipeg. However, using more than 8% of the CMA to create strata by sorting PSUs in descending order of prevalence would have had a major adverse effect on main LFS estimates for Winnipeg. The end decision was that the top 25% of PSUs[10] in terms of prevalence were isolated and stratified into twelve strata using the same optimization algorithm used to stratify PSUs outside special strata (see Section 2.5.5). That way, more strata could be created to cover the target population without as much impact on LFS estimates. Simulations were conducted to evaluate alternate strategies and this was the most favourable option.

Tables B.3 in Appendix B give the number of households in the special strata, the prevalence of the target population and the proportion of the sub-population covered by the special strata.

### 2.5.5 Stratification of the remaining PSUs

After forming special strata, which only cover a small portion of the Canadian territory, the remaining PSUs in the nine provinces other than PEI[11] are stratified within the geographic regions discussed in Section 2.5.3. To determine the number of strata that needed to be created in a region, the number of households not in special strata was divided by the targeted stratum size from Equation (2.1) in Section 2.5.2. Since this quotient is not an integer, the result was rounded up[12]. If more than one stratum was needed for a region, the PSUs were stratified first geographically and then optimally (both described below).

#### Geographic stratification

In the case of CMAs, each was divided into several pieces that would serve as the basis for stratification. The regions considered were the largest Census Subdivision (CSD), the second-largest CSD, the third-largest CSD, the remaining urban PSUs, and the rural PSUs. These regions were created only if the CMA required several strata and if the region in question met the targeted stratum size. Otherwise, they were combined with other regions.

Within a region, if only one stratum was needed then stratification is complete. If between two and nine strata were needed, the PSUs were stratified optimally (described below) within the piece. If more than ten strata were needed, the piece was first divided into super-strata – compact areas with a similar number of households – to ensure better geographic distribution of the selected PSUs in a sample. PSUs were then optimally stratified within the super-strata.

Outside CMAs, the regions were defined using the largest Census Agglomeration (CA) in an EIER, the remaining urban PSUs and the remaining rural PSUs. Within each piece, PSUs were assigned to the required number of strata using optimal stratification.

#### Optimal stratification

After geographic stratification, PSUs in regions that needed two or more strata were stratified optimally. The purpose of optimal stratification is to reduce the sampling variance of several variables of interest by grouping together PSUs with similar characteristics, creating strata that are as homogeneous as possible while conforming to the stratum size constraints determined in Section 2.5.2. This was achieved using the iterative process described below.

The algorithm used for optimal stratification is based on an iterative method developed by Friedman and Rubin (1967) and modified by Drew, Bélanger and Foy (1985). Starting with a random initial stratification with equal-sized

10. PSUs already assigned to high income strata were excluded. No Aboriginal strata were created in Winnipeg.
11. Stratification in PEI is explained in Section 2.5.6
12. In some cases this number was rounded down, usually if there were too few PSUs to create the extra stratum.

strata, the algorithm exchanges a PSU between two strata and checks whether this new stratification decreases a weighted sum of squares of auxiliary data. If the sum of squares decreased, the new stratification replaces the previous one; otherwise, the previous stratification is retained. PSU exchanges continue iteratively until no exchange leads to a decrease. The process is then repeated using different initial stratifications. The stratification associated with the smallest variance is retained[13].

The weighted sum of squares is calculated over several auxiliary characteristics. The list of these characteristics of interest (29 in total) is identical to the list from the last redesign and is available at the end of Appendix B. Household income was given three times the weight compared to the rest of the characteristics in the weighted sum of squares because income is correlated with several LFS variables. All other variables were given equal weight in the process.

After both geographic and optimal stratification are complete, the LFS geographic variable is defined by assigning unique identifiers to each created stratum and each PSU assigned to that stratum. The result is a completed LFS area frame in all provinces except for PEI. This frame is used for the first stage of sample selection (Section 2.6).

### 2.5.6 Creating strata in Prince Edward Island

For the redesign, it was decided to use one-stage sampling in PEI. As explained in Sections 2.1 and 2.2, the primary gain from two-stage sampling is to reduce the geographic spread of selected samples, thus reducing travel costs in survey collection, though with a loss in efficiency of the design. However, since PEI is a small province and many cases are now handled by telephone instead of in-person, travel costs in PEI are minimal. Historically, the other reason for two-stage sampling was the lack of a complete list of dwellings for an entire province, which limited the survey design options. However, with recent improvements to the Dwelling Universe File (Section 2.2.1), a reasonably up-to-date list of all dwellings in PEI is now available. While the CVs of monthly unemployment estimates for PEI were often far above the 7% target, it was not prudent to increase the sample size as PEI already has the highest sampling rate in the country. Therefore, it was decided to simply select dwellings in PEI from a list at random (one-stage systematic random sampling) to improve the efficiency of the design. Since CCHS needs to sample dwellings in PEI at a slightly higher rate than LFS, the ISR (Section 2.3.2) was adjusted so that each sample would contain enough dwellings for either LFS or CCHS.

Even though there are no PSUs to stratify, it was still advantageous to stratify the province for several reasons. Without strata, selecting dwellings at random could result in samples where only one dwelling is selected in one part of the island. In terms of the collection strategy, it would be hard to create a full-time workload for an interviewer covering that region and potentially very costly if that interviewer was also covering other parts of the island. In terms of sample design, that part of the island would be poorly represented in the sample and sampling variance would likely increase if that part of the province had different characteristics. Creating strata ensures that each LFS sample better represents the whole province. To stay consistent with standard geography units, geographically contiguous strata were created using Census DAs.

As explained in Section 2.2, one-sixth of the sample is replaced every month, so the PEI sample needed to be split into six rotation groups. This is more challenging without clusters in a stratum.[14] To address this, PEI geographic strata were pooled together in groups of six to form super-strata. The process controlled for the number of households in the super-strata and minimized the distance between the strata within. A modified version of the Friedman-Rubin algorithm discussed above was used. These super-strata also respected the boundaries of Charlottetown and Summerside, the two major cities in the province. Then, each stratum in a super-stratum was randomly assigned to one of six rotation groups in a way that balanced the overall number of occupied dwellings in each rotation group. Thus, one-sixth of the PEI sample can be replaced each month.

## 2.6 Sample selection strategy

Once allocation and stratification are completed, all the pieces are in place to select the sample. This section provides a conceptual description of the selection and rotation method used by the LFS in the provinces other than PEI. For PEI, systematic random sampling of dwellings within strata is used. Additional information on

13. This optimization method is known as random restart hill climbing
14. See Section 2.6.1 for how rotation groups are formed in the other provinces.

processing the growth and maintenance of the survey frame is given in Chapter 3. For a more detailed description of the sampling probabilities and the sampling weights, refer to Chapter 6.

### 2.6.1 Sample allocation of PSUs to Strata

When a two-stage design is used, survey theory stipulates that it is preferable to select the PSUs with a probability proportional to their size when this size measurement is also correlated to the estimates of interest. This is the case for the LFS. For example, the number of persons who work in a PSU is strongly correlated to the number of persons who live in the PSU. Therefore, the PSUs for the LFS are ideally selected with a probability proportional to their size. The size measure used for LFS is based on the number of households in the PSU as estimated using the DUF (explained in Section 2.2.1)[15].

The first step is to determine the number of PSUs to select in each stratum. By design, as described with determining the size of the strata, this should be six. However, due to rounding, the creation of special strata and other factors, the number of strata defined and the required sample size may not correspond to six PSUs being selected. Also, to simplify the sample rotation process – where one-sixth of the sample rotates out every month – it is preferable to select a multiple of six PSUs in each stratum. The rotation method will be discussed in detail later in the chapter.

To determine the number of PSUs to select in a stratum, the number of households to survey in the stratum is needed. Up to this point the allocation is only known at the ER-EIER-CMA intersection level. Strata are given the same sampling rate as the intersection in which they are located. Thus, within an ER-EIER-CMA intersection, the constant sampling rate implies that the sample is allocated to strata proportionally to the number of households. The household stratum allocation is given by the number of households in the stratum divided by the inverse sampling ratio (ISR) for the ER-EIER-CMA intersection.

The number of PSUs to select in the stratum is determined by the household stratum allocation divided by the target number of households to survey per selected PSU. As discussed in Section 2.5.2, this target number is six households per PSU in the Montréal, Toronto and Vancouver CMA strata, eight in the urban strata outside these three CMAs, and ten in the rural strata. If the result of the second division is closer to six PSUs than to twelve, six PSUs will be selected in the stratum. Otherwise, twelve PSUs will be selected. This can result in eighteen PSUs in rare situations.

The approach described below is based on selecting six PSUs. The same approach applies when twelve or eighteen PSUs are selected.

### 2.6.2 Overview of the RHC method

The LFS selects the PSU sample using the Rao-Hartley-Cochran (RHC) method. The RHC method is used because it allows the selection probabilities to be updated when strong growth is observed in some PSUs. The method described in Keyfitz (1951) can be combined with the RHC method to update the probabilities while maximizing the overlap of the selected PSUs before and after the update. For more information on the RHC method, see Rao, Hartley and Cochran (1962).

When using the RHC method to select multiple PSUs in a stratum, all the PSUs must first be distributed into groups, each containing roughly the same number of PSUs – plus or minus one. In the case of the LFS, the groups used are the six rotation groups. After the PSUs have been distributed to the rotation groups, one PSU is selected per group with probability proportional to size within the group. This can be summarized by the following equation:

$$\pi_{hij} = \frac{M_{hij}}{\sum_{j \in hi} M_{hij}} \qquad (2.2)$$

where

$M_{hij}$ is the number of households in PSU *j* in rotation group *i* of stratum *h*.

15. In practice, the size measurement used is the inverse sampling ratio, which is derived from the number of households. More information on this calculation is provided in Section 2.7.1.

$\sum_{j \in hi} M_{hij}$ is the total number of households in all the PSUs in rotation group *i* of stratum *h*.

$\pi_{hij}$ is the selection probability of PSU *j* in rotation group *i* of stratum *h*.

Rather than using the number of households, the LFS uses the rounded inverse sampling ratio of the PSU $\left( ISR^*_{hij} \right)$ as a size measure for the PSU – described below. These values are used mainly because of the way the sample is selected in the second stage. It will be shown later that this is not an extreme departure from using the number of households in terms of sampling probability for the PSU.

**Second-stage selection probabilities**

Dwellings are selected from within the selected PSUs with probabilities that ensure that all households in the stratum have the same overall probability of selection. This is often referred to as a self-weighted design.

At the second-stage, dwellings are selected from the PSU listing line generated by the Address Register and/or field listing of the PSU[16] using systematic sampling where households are selected at regular intervals. This method is recommended because it is simple to use, ensures a good distribution of the households selected in the PSU, controls the overlap of samples and facilitates adding new dwellings to the PSU. To select the systematic sample of dwellings, the PSU ISR, $ISR_{hij}$, and a starting point on the list must be determined. $ISR_{hij}$ can be obtained from the number of households in the PSU and the ISR of the stratum, $ISR_h$, using the following equation:

$$ISR_{hij} = \left( \frac{M_{hij}}{\sum_{j \in hi} M_{hij}} \right) ISR_h \qquad (2.3)$$

where

$ISR_{hij}$ is the inverse sampling ratio in PSU *j* in rotation group *i* of stratum *h*.

$ISR_h$ is the inverse sampling ratio of stratum *h* established during the allocation of the sample.

Since $ISR_h$ is constant for all the PSUs of a group, $ISR_{hij}$ is proportional to the number of households in the PSU. For the redesign, a switch to simple random sampling was considered. However, a study showed that adjacent dwellings have correlated responses. This implies that for monthly estimates, systematic sampling should lead to a better representation of the PSU since the sample is guaranteed to be spread out over the entire region. Also, systematic sampling can give reduced variance of estimates of month-to-month change as neighbouring households are rotated in to replace those who are rotated out.

The LFS selection system cannot use these *ISR*s directly and instead is configured to use integer inverse sampling ratios *ISR**. The result of Equation 2.3 is therefore rounded up or down so that

$$\sum_{j \in hi} ISR^*_{hij} = ISR^*_{hi} = ISR^*_h, \forall i \in h \qquad (2.4)$$

This is called controlled rounding. The second-stage selection probability of a household when PSU *j* in rotation group *i* of stratum *h* is selected is $1 / ISR^*_{hij}$.

The rounded value $ISR^*_{hij}$ has some useful interpretations. First, it is the sampling interval to use in systematic sampling if the corresponding PSU is selected in the first stage. By applying this sampling interval, the appropriate number of households will be selected in the PSU[17]. Second, $ISR^*_{hij}$ is the number of distinct samples available in the PSU. In LFS terminology, this concept is called the number of *random starts*.

16. Chapter 3 describes how the Address Register and field listing are used to create the sample frame.
17. This target corresponds to the number of households in the group divided by the stratum inverse sampling ratio.

### First-stage selection probabilities

As stated earlier, a self-weighted design is achievable when the first and second stage probabilities are in agreement. The second stage probabilities were defined earlier as $1/ISR^*_{hij}$ . In order to preserve the self-weighting aspect, the values $ISR^*_{hij}$ must be used as size values in the probability proportional to size sample. The first-stage selection probability associated with each PSU is therefore:

$$\pi^*_{hij} = \frac{ISR^*_{hij}}{\sum_{j \in hi} ISR^*_{hij}} = \frac{ISR^*_{hij}}{ISR^*_h}. \qquad (2.5)$$

This is not an extreme change from using the number of households as a size measure. As stated earlier, $ISR_{hij}$ is proportional to the number of households in the PSU. In this case,

$$\pi^*_{hij} = \frac{ISR^*_{hij}}{\sum_{j \in hi} ISR^*_{hij}} \approx \frac{M_{hij}}{\sum_{j \in hi} M_{hij}} = \pi_{hij}. \qquad (2.6)$$

The only difference between these two probabilities is due to the controlled rounding of $ISR^*_{hij}$ . As a result, the overall selection probability of household *k* in PSU *j* in rotation group *i* of stratum *h* is:

$$\pi^*_{hijk} = \frac{ISR^*_{hij}}{\sum_{j \in hi} ISR^*_{hij}} \times \frac{1}{ISR^*_{hij}} = \frac{1}{ISR^*_h}. \qquad (2.7)$$

As required, Equation 2.7 suggests that the selection probability is the same for all households in the same stratum. The LFS sample design is therefore self-weighted within the stratum.

## 2.6.3 PSU and start selection

In practice, to select a PSU in a rotation group, the PSUs of a rotation group are put in random order. A random whole number *U* is then drawn from a uniform distribution on the interval $\left[1, ISR^*_h\right]$. This random number *U* has two functions. First, it is used to identify the first PSU selected. This PSU is the first for which the cumulative total of the $ISR^*_{hij}$ is greater than or equal to *U* (or $\sum_{j \le j^*} ISR^*_{hij} \ge U$ where the indicator *j* follows the random order).

It also determines the number of random starts to use in this first PSU before moving on to the next PSU. The number of starts to use in the first PSU is $D_{j^*} = \left(\sum_{j \le j^*} ISR^*_{hij}\right) - U + 1$. Lastly, a second random whole number $U_{j^*} \in \left[1, ISR^*_{hij^*}\right]$ [18] is selected. This number indicates the first random start to use to select the sample of dwellings for the PSU *j**. The systematic sample for a selected PSU *hij* is composed of the dwelling whose line number on the dwelling frame is equal to the starting point $U_{j^*}$ , and of other dwellings whose additional lines are in intervals of $ISR^*_{hij}$ . Therefore, dwellings are selected with line numbers such that $d = U_{j^*} + t \times ISR^*_{hij}$ , *t* = 0,1,2,… until *d* exceeds the number of lines available.

These dwellings will remain in the sample for a period of six months.

18. This second random number has two functions. It takes into account the fact that the sample size associated with the last random start is sometimes smaller than that of the first starts. We therefore hope to stabilize the global sample size over time. It also lays the groundwork for applying the rule of the minimum number of starts to use.

Gray (1973) and Alexander, Ernst and Haas (1982) use two different approaches to illustrate that this method produces a sample that respects the selection probabilities specified. Laflamme (2003) demonstrates the sample selection process using a diagram.

### 2.6.4 Sample rotation

Section 2.6.3 describes how the first sample of dwellings was selected in each group created using the RHC method. After a period of six months, it is necessary to replace this sample with new dwellings. By continuing with the example given at the end of the previous section, the first sample corresponded to the random start $U_{j^*}$ of the PSU *j**.

If the number of random starts to use from PSU *j** is $D_{j^*} = 1$, the second sample of dwellings will correspond to the start $U_{j^*+1}$ of the next PSU, *j**+1, where $U_{j^*+1} \in \left[1, ISR^*_{hi(j^*+1)}\right]$. Otherwise, if $D_{j^*} > 1$, the second sample will correspond to the start $U_{j^*} + 1$ of PSU *j** (*i.e.*, the neighbours of the previous sample). If $U_{j^*} + \ > ISR_{hij^*}$, the selection loops back to use start 1 of PSU *j**. Generally speaking, with this method, PSU *j* remains in the sample for $D_j$ periods of six months. When it is necessary to replace the surveyed dwellings, the next random start is used. After $D_j$ periods, the sample moves to the value $U_{j+1}$ of PSU *j*+1. This PSU will remain in the sample until all its random starts have been used. The same goes for the PSUs that are added to the sample at a later date.

This method produces the expected results: the selection probabilities are always respected over time. Unfortunately, it has a major inconvenience. As discussed, the first PSU remains in the sample for a random number of periods, and sometimes this number is small. This rapid rotation of the first PSU selected would lead to an inefficient use of the survey's limited resources. In fact, adding a PSU to a sample requires a great deal of work, including preparing the material, possibly listing the PSU and sometimes hiring and training an interviewer. To be effective, it would be preferable to amortize this investment by avoiding a too-rapid rotation of the first PSU as much as possible.

To overcome this problem, the LFS developed a correction that increases the number of random starts to use from the first PSU without introducing a bias into the selection probabilities. When $D_{j^*}$ is too small, based on a pre-determined criterion, it is increased in order to keep this PSU in the sample longer. In this case, the number of starts to use for PSU *j**+1 must be reduced proportionally in order to avoid introducing a bias into the selection probabilities. Some constraints are required to ensure that the increase in the number of starts associated with the first PSU will not reduce the number of starts to survey from the second PSU by too much. Gray (1973) shows that this approach does not bias the selection probabilities, while Laflamme (2003) provides explanations on these constraints.

This method is applied separately to each rotation group. However, the samples are not all rotated at the same time. An RHC group in rotation group 1 is rotated in January and July of every year. The RHC groups in rotation group 2 are rotated in February and August, and so on. By using this method, at the start of the redesign, a list can be produced containing all the starts that will be in the LFS sample for each month over the next ten years.

**Figure 2.1**
**LFS Sample Rotation**

| Rotation Group | Survey Month | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC |
| 2 | 6th | | | | | | | | | | | |
| 3 | 5th | 6th | | | | | | | | | | |
| 4 | 4th | 5th | 6th | | | | | | | | | |
| 5 | 3rd | 4th | 5th | 6th | | | | | | | | |
| 6 | 2nd | 3rd | 4th | 5th | 6th | | | | | | | |
| 1 | 1st | 2nd | 3rd | 4th | 5th | 6th | | | | | | |
| 2 | | 1st | 2nd | 3rd | 4th | 5th | 6th | | | | | |
| 3 | | | 1st | 2nd | 3rd | 4th | 5th | 6th | | | | |
| 4 | | | | 1st | 2nd | 3rd | 4th | 5th | 6th | | | |
| 5 | | | | | 1st | 2nd | 3rd | 4th | 5th | 6th | | |
| 6 | | | | | | 1st | 2nd | 3rd | 4th | 5th | 6th | |
| 1 | | | | | | | 1st | 2nd | 3rd | 4th | 5th | 6th |
| 2 | | | | | | | | 1st | 2nd | 3rd | 4th | 5th |
| 3 | | | | | | | | | 1st | 2nd | 3rd | 4th |
| 4 | | | | | | | | | | 1st | 2nd | 3rd |
| 5 | | | | | | | | | | | 1st | 2nd |
| 6 | | | | | | | | | | | | 1st |

This diagram illustrates the LFS sample rotation design. The colors indicate which rotation group the dwellings belong to (Orange is rotation group 1, pink is rotation group 2, blue is rotation group 3, green is rotation group 4, grey is rotation group 5 and yellow is rotation group 6). The numbers in the boxes indicate the number of months that the dwellings associated with a given rotation group have been part of the survey. As shown by the diagram, one-sixth of the sample is renewed monthly. So, say in April, dwellings from the blue rotation group are in their second month of the survey, while dwellings from the grey rotation panel are in their sixth and last month of participation. Dwellings from the grey rotation panel will be replaced by new dwellings in May, as seen in the diagram. Dwellings that rotate-out are generally replaced by dwellings from the same respective primary sampling units.

# Chapter 3 Dwelling frame creation and maintenance

## 3.0 Introduction

As described in the previous chapter, the Labour Force Survey (LFS) uses a two-stage sample design in all provinces except for Prince Edward Island. An advantage of this approach is that the sample is concentrated in a limited number of areas; therefore, it is possible to conduct personal interviews. At the first stage, primary sampling units (PSUs) – also called clusters – corresponding to geographic areas are selected. These are relatively small parcels of land, often Census Dissemination Areas (DA). Within the selected PSUs, dwellings are selected at the second stage of sampling.

At both stages of the sampling process, a survey frame, i.e., a list of all the units (clusters or dwellings) that are part of the target population, is required. A good quality frame will have limited coverage errors and facilitate contact with the sampled units. Given that new units are continually being added and removed from the target population, it is important that maintenance and updates are performed on the sampling frame. Details of the frame creation for PSUs and the design aspects for the selection of households were described in Chapter 2.

## 3.1 Dwelling frame creation

Within selected PSUs, a complete list of dwellings (a frame) is required in order to select the second-stage sample. The list is obtained either through a listing exercise performed in the field or from an existing list, specifically the Address Register (AR). Once the dwelling list is available, it will be used as long as the PSU is in sample. One continuing challenge is to determine which newly sampled PSUs should undergo listing and which can rely on the AR information. Field listing is a more costly option that should be avoided whenever possible. It usually occurs when the information on existing lists is of low quality.

### 3.1.1 The Address Register

The AR is a database that was initially created for the 1991 Canadian Census of Population, with the purpose of improving census coverage. It was created using several administrative files, such as telephone billing files and building permit files. Immediately after that census, the AR was updated using the list of addresses created during the census enumeration process. Since that first iteration, the AR has continued to be updated quarterly using administrative files and the census listing program, and census information available every five years.

The AR was originally designed to provide and maintain a list of addresses for communities with a population over 50,000. The coverage of the AR was expanded following each subsequent census to include smaller population centres and regions outside population centres. Currently, the AR has national coverage, though it is more accurate in population centres.

In 2015, the AR included over 15 million addresses. The vast majority of these addresses – about 90% – were found to be valid residential dwellings during the 2011 Census. Of the remaining addresses, 7% were obtained through updates from administrative files and field listing in preparation for Census 2016 and 3% were valid dwellings during a previous census.

To appear on the AR, a residential dwelling must possess a valid standard civic address or any sort of descriptive address. For survey purposes, the descriptive addresses are often incomplete and may not provide enough information to locate the dwelling. Where there is a substantial proportion of descriptive addresses, the area may have to be listed in the field.

Two key files are extracted from the AR database for the LFS dwelling frame creation process: the Dwelling Universe File and the Residential Telephone File.

#### Dwelling Universe File

The Dwelling Universe File (DUF) is an extraction of addresses from the AR. Rules are applied to ensure that the list only contains dwellings that correspond to the target population of the LFS. These rules evolve over time as methods to detect spurious or duplicate addresses improve. Collective dwellings are also a small part of the LFS target population and these dwellings are available through the AR extraction process.

#### Residential Telephone File

The Residential Telephone File (RTF) is a list of residential telephone numbers valid in Canada. Many of them (88% in 2015) can be associated with a dwelling address found on the DUF. The RTF can therefore be used to add telephone numbers – key contact information – to a large portion of sampled dwellings.

### 3.1.2 The National Geographic Database

To use the AR in a two-stage design context, each address must first be assigned to a specific PSU. This is achieved by linking the AR to the National Geographic Database (NGD). The NGD contains map layers that include PSU boundaries, street networks, waterways, and other geographical markers. This information can be used to link addresses to street sections. These sections can be at the block level (a block is a polygon with street segment sides contained within a DA) or more precisely at the block-face level (a single street segment). These sections are then associated with a DA or PSU which effectively associates a dwelling with the PSU.

The NGD is managed in partnership with Elections Canada and is constantly changing due to the regular addition of roads and geographic boundary updates such as municipal boundaries. Every three months a new vintage of the NGD is released.

### 3.1.3 Ordering the list of addresses

The addresses on the dwelling frame must be organized into a list with a specific order that can be maintained over time. This ordering helps to facilitate finding selected dwellings and can help interviewers to recognize any list omissions. The ordering of the addresses is created by a sequencing algorithm which lists the block-faces in an order that covers the entire PSU while minimizing the total distance travelled by the interviewer when verifying the list of addresses. This algorithm uses the geographical information within the PSU coming from the NGD and is most helpful to field staff when all addresses can be block-face geocoded. The algorithm is run for the entire frame of PSUs for each vintage of the NGD. This means that in each selected PSU, the list of addresses is put in a specific order to facilitate and optimize listing.

## 3.2 Loading and Field Listing

Once the dwellings have been assigned to their PSUs, quality indicators for the list of addresses can be developed. The quality determines if the region will require field listing or if the AR-NGD information will suffice as the list to be used as the sampling frame for dwellings in the PSU.

Ideally, the lists in all PSUs would be verified in the field (field listed), but the budget restricts the number of PSUs that can fall in this category. The quality of the list of addresses for a given PSU depends on the quality of the AR, the quality of the NGD, and the effectiveness of the DUF eligibility rules. The goal of this strategy is to make as much use of the AR as possible while at the same time taking into account the fact that its quality varies for different regions.

The AR quality is known to be highest in population centres. These population centres largely correspond to the “mail-out area”, where the census collection method is to reach households by mail. This area corresponds to about 80% of dwellings. Based on this information, PSUs are classified into one of three groups:

**AR Group 0** are PSUs in the mail-out area. No initial listing in the field is performed and the first sample of units is selected from the AR-based list. While not field listed “as an LFS PSU”, a substantial portion of the mail-out area undergoes field listing under the census listing program. The results of that listing are processed by the AR team and ultimately appear on the DUF.

**AR Group 1** are non-mail-out PSUs with no initial listing. The AR list is assessed to be of good quality, based on a collection of statistics and indicators. The initial sample of dwellings is selected directly from the AR-based list.

**AR Group 2** are non-mail-out PSUs with initial listing. The PSU must be field listed before the first sample selection occurs.

The 2015 allocation assigned 72% of the sampled PSUs to AR Group 0, 19% to AR Group 1 and 9% to AR Group 2. This is a major change from the launch of the 2005 design where fully 61% of PSUs required initial listing. As PSUs rotate in and out of the sample and as the quality of the AR evolves (especially after the 2016 Census) the distribution of the AR Groups will likely change.

### 3.2.1 Initial loading

For AR Group 0 or AR Group 1 PSUs, the dwelling list used in sample selection is populated from the list of dwelling addresses available on the DUF linked to these PSUs. This process is called initial loading. The LFS sample of dwellings is selected directly from this list.

Unlisted PSUs tend to have a higher proportion of sampled units coded “invalid” or “demolished” at the time of survey collection. Coverage errors will be discussed in Chapter 8.

### 3.2.2 Initial listing

The PSUs in AR Group 2 must undergo initial listing. The goal of initial listing is to prepare a complete and accurate dwelling list for the first sample selection in a PSU. The initial listing case is pre-filled with the dwellings associated with that PSU according to the DUF. Each dwelling in the list is validated, modified or deactivated by field staff. New dwellings can also be added to the list.

### PSU mapping

In order to complete the field listing effectively, the PSU boundaries must be displayed on a map. Proper translation of the map contents in relation to physical features on the ground is paramount in determining which dwellings belong to the PSU. Further, the block numbers and address ranges on the map can help pinpoint specific addresses. Dwelling addresses or descriptions are captured by the field interviewer using the Statistics Canada Listing Application. PSU maps are generated using Generalized Mapping System software in place since 2009. Appendix D contains examples of PSU maps and describes more details about their creation and uses.

#### Listing collectives

The listing of collectives is not as clear-cut as with privately-occupied dwellings. There are two main criteria for listing collectives. First, inmates of institutions are not part of the population covered by the LFS. Likewise, temporary residents with a usual place of residence elsewhere are not eligible. Generally only the owners’ residence, any staff residences, and dwellings for non-institutionalized residents (*e.g.*, units in a seniors’ residence) would be listed.

## 3.3 Frame maintenance

Regardless of whether or not the PSU underwent initial listing, each month there is an opportunity to update or correct the dwelling list. Therefore, most frame problems are temporary and can be rectified for subsequent sampling occasions.

### 3.3.1 List update and list maintenance

Once a PSU has been selected, regular updates to the address list can come either quarterly from each new vintage of the DUF (list update) or monthly from field verification (list maintenance). For AR Group 0 clusters, a combination of list update and list maintenance is used. For AR Group 1 and 2 clusters, list maintenance is the main source.

In list maintenance, dwellings can be added, modified or deactivated (with some reason for the deactivation). Dwellings can be moved in the listing order, affecting a print sequence number, but the permanent within cluster ID number, the listing line, remains fixed. This approach allows the interviewer to have a preferred listing order while effectively preserving the sample history of each dwelling.

Maintenance is normally conducted "on rotation", meaning during the first month of sampling (*e.g.*, January or July birth months for rotation 1 PSUs). Typically, the interviewer must visit the PSU because at least some of the newly selected dwellings need to be contacted in person.

#### Interviewer Selected Dwellings

List maintenance can trigger Interviewer Selected Dwellings (ISDs). These are new LFS cases for the CAPI interviewer to complete.

There are two forms of ISDs created during list maintenance. First, during the life of the PSU, the interviewer can add new dwellings on a regular basis as the population grows. Since the dwelling list is open-ended, additional dwellings can be selected in the field. The structures added to the end of the list are sampled using the PSU-level inverse sampling rate (ISR) and next-line-to-be-interviewed provided from the latest sample selection in the PSU. Once a dwelling is selected, the next-line-to-be-interviewed is incremented by the PSU ISR.

The second form of ISDs is known as multiples. During the process of interviewing within a selected dwelling, the interviewer may determine that separate dwellings exist within the structure that were not identified in the list. Typically these are basement or upper units not evident from the street. Since the dwelling list does not contain the extra units as separate lines, these dwellings have no probability of being selected over the lifetime of the PSU. To compensate for the missed dwellings in this and any other similar unresolved cases, all missed units are selected to be in sample along with the original dwelling. They are added to the list as multiples of the originally selected dwelling and the application generates a case for each multiple.

### 3.3.2 Treatment of growth areas

Since PSU dwelling lists are open-ended, there is potential for extreme growth. Interviewers may not be able to maintain large lists because of the cost associated with this maintenance, and the time required to conduct interviews for the large influx of new sample that such a large list implies. Although this extreme growth is observed in less than 1% of PSUs, options must be available to manage and treat it.

#### PSU sub-sampling

Based on feedback from the field, the PSUs with large growth may hinder the ability of the interviewer to complete all the assigned interviews. This can be even more difficult during a birth assignment, especially if the fraction of households requiring in-person interviews is high. In such cases of isolated growth, the PSU is sub-sampled to reduce the burden. The LFS uses two forms of sub-sampling.

The first is a simple modification to the sampling rate for the specific PSU. This technique – also called cluster or mechanical sub-sampling – is used for the majority of cases. Often, it is sufficient to decrease the sampling rate by a factor of two in order to reduce the interviewer's workload by half.

The second form of sub-sampling is the insertion of an additional stage of sample selection. In this technique, sub-clusters are formed as second-stage units (SSUs). By convention, PSUs can be referred to as clusters and parts of PSUs as sub-clusters. In cases of large growth, head office staff delineate four or more sub-clusters of approximately equal size in terms of number of households within the PSU. Two of the SSUs are then selected for survey activity and sub-sampling factors are created.

Sub-sampling modifications affect the sampling probability of the households. Descriptions of the adjustments to account for this are found in the explanation of the weights in Chapter 6.

#### Stratum update

On rare occasions, the growth in a PSU is so extreme it causes a more than tenfold increase in the number of households. In this scenario, PSU sub-sampling may introduce extreme sampling factors or be insufficient to reduce the interviewer's workload. In addition, the sub-sampling factors can create high variability amongst the sampling probabilities and may affect the precision of estimates. In such cases, it is better to redesign the stratum. Typically, other PSUs in the stratum will also have exhibited significant growth.

For a stratum level redesign, the original PSUs exhibiting extreme growth are re-delineated into several new PSUs having approximately 230 households each, which is the average size of a PSU. Estimated household counts for all PSUs in the stratum are required, whether they are newly formed or retaining their original boundary. These counts can often be derived based on the latest DUF. With these revised inputs, the stratum update program is run to re-form the random rotation groups and re-establish the PSU level sampling fractions. This program, based on Keyfitz (1951), as modified by Drew, Choudhry, and Gray (1978), retains as many of the selected PSUs as possible at the time of the update.

The newly selected PSUs must be field listed or loaded with information from the AR. The new sample is phased-in over six months.

### 3.3.3 Monitoring PSU yield

Through time, the PSU yield of households is carefully monitored. PSU with exceptionally small or large household yields may need special attention or treatment. A very low household yield suggests a fundamental change since the design counts were established in June 2013. A large household yield usually indicates areas of growth, but may also indicate dwellings shifted into the wrong PSU on the DUF. Field follow-up or head office investigations are done to justify or correct discrepancies.

### 3.3.4 Sample size stabilization

Over time there is a slow increase in the size of the population. Left unchecked, this growth would increase the sample size and survey collection costs. In order to keep the sample size under control, stabilization can be used.

#### Unit targets

The first step of stabilization is to determine where stabilization is necessary. Unit targets – the number of sampled units required in a region to obtain the desired sample of households – are determined. The unit targets take into account that some units on the frame will not necessarily be valid dwellings and that some fraction of valid dwellings are not occupied (*i.e.*, not households). Each stabilization area is a collection of strata that roughly corresponds with an Employment Insurance Economic Region (EIER) or some portion of an EIER. Unit targets should function for all rotations and rarely require updating. The results of recent collection generally indicate where adjustments to the unit targets are warranted – either due to an observed household shortfall or surplus.

#### Stabilization selection

The unit targets are compared with the number of units obtained from sampling from the most up–to-date dwelling lists at the prescribed rates. Regions requiring stabilization are those where the sample obtained from the most recent dwelling lists contains more units than required according to the unit targets. The number of units to drop is the number of units in the initial sample minus the unit target.

Some areas are defined with no expectation to drop units in that area because of the small sample size and high relative variability. Large growth PSUs with sub-sampling are also exempt from stabilization to avoid further inflation to the sub-sampling factors that are already present. From the remaining units, a systematic subsample of units is selected to drop from the collection process. The selection probabilities of the units not dropped are adjusted to ensure a proper representation of the population.

Other surveys that select units from the LFS frame can do their own stabilization, dropping units from their initial sample. These surveys are discussed in Chapter 9.

#### Stabilization weight adjustment

The stabilization weight, used to compensate for dwellings dropped from the sample, is calculated after the drop is completed. Not all strata in one stabilization area have the same stratum ISR and the calculation of the weight adjustment takes this into account, ensuring that the sampled units properly represent the population.

The following example illustrates how the stabilization factors do not affect the weighted contribution from the entire stabilization area.

Imagine a stabilization area of three strata, A, B, and C with stratum level ISRs of 400, 500 and 600 and pre-stabilization unit yields of 10, 10 and 10 respectively. Further assume the unit target for this stabilization area is 28, meaning two units should be dropped. For this example, it is assumed that one unit was dropped from stratum A and one unit was dropped from stratum B.

The weighted contribution from this stabilization area should be 15,000 = 10x400+10x500+10x600. With the two units dropped, the weighted contribution becomes 9x400+9x500+10x600=14,100. The stabilization factor is such that the weighted contribution from this stabilization area is preserved. In this example, the factor of 15,000/14,100 is applied to the selected units that remain in sample and the contribution of these units to this stabilization area is exactly 15,000 with this adjustment applied to the weights.

### Special considerations

Dwellings selected in the field due to growth in the PSU are identified after the stabilization process and therefore have no chance to be included in the stabilization program. In theory, these dwellings should not have a stabilization weight applied. However, our current systems assign the stabilization factors at the stratum level, and any ISDs are subject to the same stabilization factor as other units in the stratum. The impact is minimal as the number of growth ISDs is small and the stabilization factors are close to or exactly 1. Multiples, multi-unit dwellings misidentified as single residences[19], are given the stabilization weight, in effect appropriating the weight of the main residence.

19. See Section 3.3.1 for more information.

# Chapter 4 Collection

## 4.0 Introduction

Since the Labour Force Survey (LFS) is a monthly survey, the data used to produce the various LFS estimates is obtained by contacting the sampled households each month.

The complete LFS operations schedule for a given month involves four different phases (pre-processing, collection, processing and dissemination) which last about four full weeks in total. Therefore, the collection activities must follow a strict timetable established in accordance with the requirements of the other survey processes.

Data collection for the LFS takes place during the week that follows the LFS reference week, which is usually the week containing the fifteenth day of that month. Interviews begin on the Sunday of the collection week and generally continue until Tuesday of the following week. December is an exception: to avoid conducting interviews too close to Christmas, collection is done earlier than in other months. On rare occasions, collection is extended by one day.

A team of roughly 1,300 interviewers is involved each month in collection activities. The data collected by the interviewers is transmitted to head office for processing.

This chapter describes the collection methods used and some specific survey rules related to collection.

## 4.1 Collection methods

For collection purposes, there are two types of households in the LFS: “births” and “subsequents”. “Births” are households that are in their first month of participation in the survey. They represent about one-sixth of the monthly sample. “Subsequents” are households that are between their second and sixth month of participation in the survey. They represent about five-sixths of the monthly sample. It should be noted that, with respect to collection, births do not always correspond to the first month of survey collection for the dwelling. In the event of a complete change in the household occupants (*e.g.*, due to moving), nonresponse, and/or dwelling vacancy in previous months, a household is still considered a “birth” since it is the first month of response for those occupants.

Since March 2015, LFS interviews are now conducted using three collection methods: computer-assisted personal interviewing (CAPI), computer-assisted telephone interviewing (CATI) and computer-assisted web interviewing (CAWI), also referred to as Electronic Questionnaire (EQ). From March to October 2015, EQ was offered as an option to half of each rotation group. The EQ option is now offered to all eligible respondents beginning in November 2015. The collection method used for a specific case depends on several criteria linked to the type of household (birth or subsequent), the household’s eligibility for EQ and the household respondent’s preferences. Figure 4.1 at the end of this chapter provides a visual depiction of the collection method assignment process.

### 4.1.1 Collection methods for births

Before 2004, CAPI was used for all households in their first month of the survey, with interviewers visiting in person to conduct the interview.

CATI interviews for births were introduced in 2004 to reduce the collection costs associated with an initial personal interview. This approach, called Telephone First Contact (TFC), uses the Residential Telephone File (RTF)[20] to obtain a telephone number for selected households. TFC is used in areas where the information used to create the RTF is updated more regularly and where the dwelling addresses tend to have a standard form, which provides a better match between addresses and telephone numbers.

In 2015, the list of TFC strata was extended and now covers 977 of the 1,153 strata in the ten provinces. Of all households in these strata for which a telephone number is obtained, only about 3,000 (28% of births) are assigned to the CATI call-centers for CATI interviewing. This selection of the ‘most CATI-suitable’ households

20. See Chapter 3 for more information on the RTF.

is based on quality indicators derived from the characteristics of the address of the household and from RTF variables. Fixing the number of households assigned to CATI each month to 3,000 helps the regional offices with respect to planning and resource management.

The TFC approach has also been extended to allow CAPI interviewers to use the landline telephone numbers linked with selected households that were not retained for CATI interviewing. This extension covers both TFC strata and non-TFC strata households. CAPI interviewers may choose, since September 2012, to establish a first contact by telephone. On a monthly basis, a telephone number is provided for about half of the households assigned to CAPI interviewers. Of those, a first contact attempt by telephone is usually made for about 60% of households, and about 25% do not require any CAPI in-person visits.

EQ is not currently available for births as there is currently no administrative list of email addresses available for selected households.

### 4.1.2 Collection methods for subsequents

At the end of the birth interview, a series of questions is used to determine the household's eligibility for EQ and the household's preferred collection method for subsequent months.

In order to be eligible for EQ, a number of conditions must be met. First, the birth interview must be considered complete. Next, the person identified as the best contact person for subsequent months must be the person the interviewer is talking to and the person who provided the information for the birth month for all the members of the household. Fictitious names must not have been provided for any of the household members. Finally, a valid ten-digit telephone number and a complete listing address must also have been recorded.

If all conditions are met, the respondent is asked if he or she would prefer to complete the survey on the internet next month or to have a Statistics Canada interviewer contact him or her directly. If internet is the preferred option, the respondent is asked to provide an email address. If an email address is provided, this household is assigned to EQ for the remaining months.

If not all the conditions are met, or if the respondent refuses to provide an email address, then an offer to proceed with CATI interviewing for next month is made. If the respondent does not have a telephone number or indicates a preference for in-person visit, then CAPI interviewing takes place the next month.

For households answering by CAPI or CATI for a subsequent month, eligibility for EQ is re-assessed on a monthly basis. If conditions are met, an offer to proceed with EQ for the next month is made. Similarly, for households answering by CAPI for a subsequent month that are not eligible for EQ, an offer to proceed with CATI for the next month is made.

Before the introduction of EQ, about 96% of subsequent interviews were conducted by CATI. With the full implementation of EQ, about 20% of subsequent interviews will be conducted by EQ, about 76% by CATI and the remaining 4% by CAPI.

### 4.1.3 Collection mode transfers

The LFS has some capacity to transfer a case from one collection mode to another. The following are the four situations that will trigger such a transfer.

a. The contact component of the questionnaire is used to verify the address of the household. This is essential to ensure that the household contacted lives in the selected dwelling. If a confirmation is obtained that either the telephone number is invalid or that it does not lead to the selected dwelling, additional sources are searched to find a valid telephone number for the selected dwelling. If a valid number cannot be found, the CATI case will be transferred to a field interviewer who will go to the selected dwelling for CAPI interviewing.
b. When a telephone number leads to no contact for two consecutive months, even if the number is not confirmed as invalid, the case will be transferred to a CAPI interviewer for its third month in the survey.
c. If the respondent requests a personal interview, CATI cases can be transferred to field interviewers during collection.

d. If no response has been received from a household assigned to EQ within the first four days of collection, then the case will be transferred to a CATI interviewer who will use the provided telephone number to attempt to conduct a telephone interview.

## 4.2 Survey rules related to collection

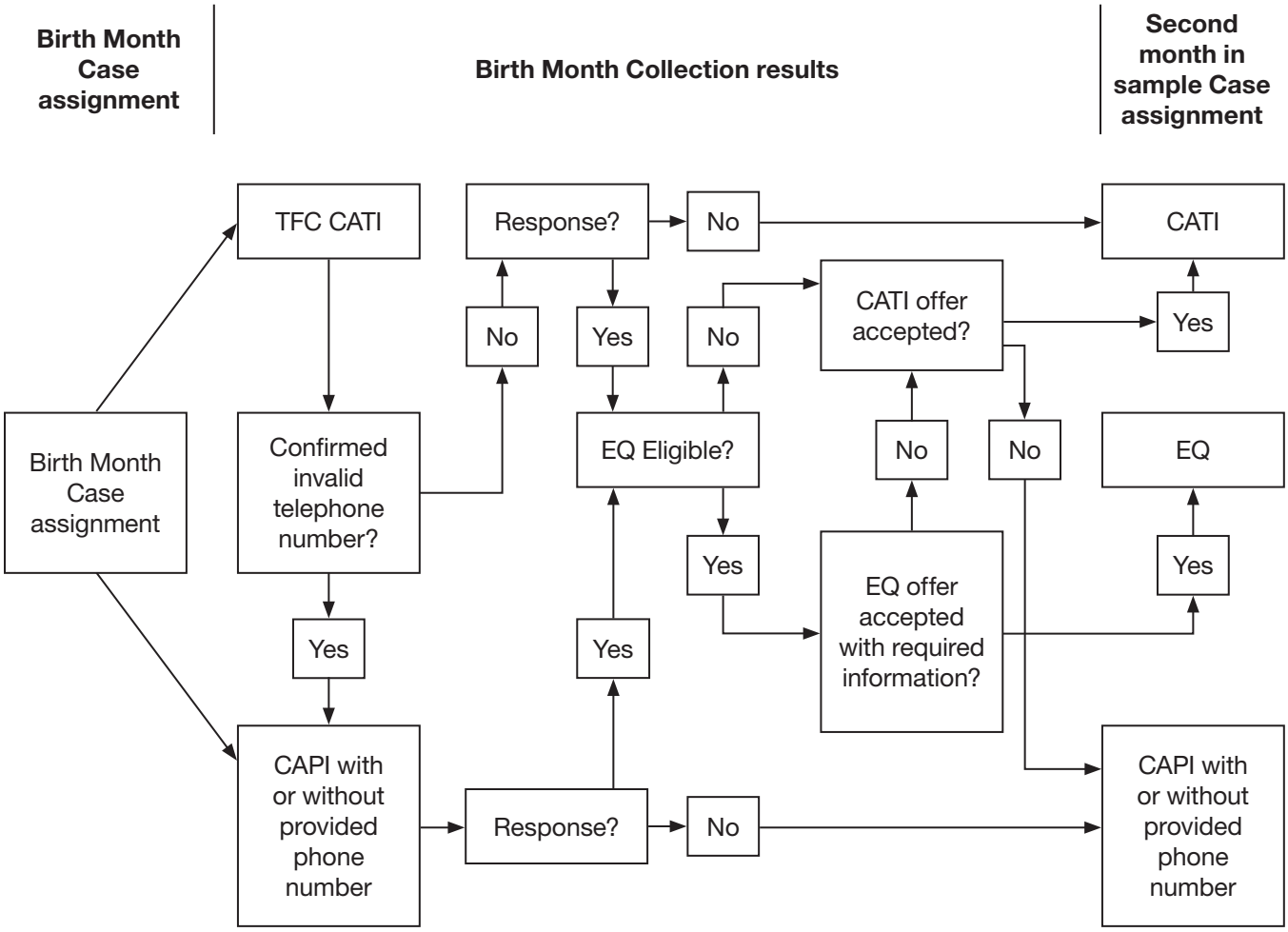
The LFS has several collection rules to reduce respondent burden and collection costs, while still achieving high response rates and data quality. The most significant ones are listed below.

Responses for household members are usually provided by a single, well-informed member of the household. Responses for household members obtained indirectly from the respondent is called “proxy response”. It is allowed because it would be too time-consuming and costly to make several visits or calls to obtain the information directly from each household member. Approximately 60% of the data in the LFS is obtained using this method.

During the birth interview, the interviewer collects socio-demographic and labour force information for all members of the selected household. In subsequent interviews, the interviewer will verify the list of household members, then collect current month labour information. For persons aged 70 years or over, the burden imposed on the respondent is reduced by reusing the responses on labour information provided in the birth interview for subsequent months.

Different types of letters are sent to the selected households to help maintaining a high response rate. For example, when a household is selected for the first time, an introductory letter and information brochure are mailed out prior to the first interview. Refusal letters are also sent out to convince reluctant households to participate.

**Figure 4.1 Flowchart of collection method assignments for births and second month in sample**



# Chapter 5 Processing and imputation

## 5.0 Introduction

After collection, the Labour Force Survey (LFS) data go through several steps of processing before estimates are produced. To facilitate production of estimates from a complete and consistent microdata file, editing, imputation, and weight adjustments are used to identify and compensate for invalid, inconsistent, and missing data. Data processing can be divided into the following four steps:

1. Receipt of data from the regional offices and Phase I editing
2. Phase II editing
3. Hot-deck imputation
4. Post-imputation processing

Invalid and inconsistent data is identified and replaced with valid, consistent data using edits and various imputation methods, depending on the type of nonresponse. Item nonresponse, where one or more questionnaire items is unknown, is treated by carry-forward imputation, imputation by deduction, or hot-deck imputation, depending on the response history of the respondent and what survey data was collected for the record. Person nonresponse, where it is not possible to obtain any survey information for a person, is treated by hot-deck imputation. Household nonresponse, where it is not possible to obtain survey information for the entire household, is treated by hot-deck imputation or a nonresponse weight adjustment (see Chapter 6), depending on the response history of the household.

The following sections will explain the processing steps in more detail, with much of the focus on the hot-deck imputation system.

## 5.1 Receipt of data and Phase I editing

During the collection period, cases with completed interviews are transmitted from the regional offices to the head office on a daily basis. Data are then processed at the head office. The LFS collects socio-demographic (*e.g.*, age, sex, education, immigration status, and aboriginal status) and labour force (*e.g.*, labour force status, class of worker, industry, and earnings) data. Item/block editing and consistency editing are performed in several stages. Phase I editing includes four stages: record acceptance, demographic editing, Labour Force Information (LFI) item acceptance, and industry/occupation editing.

In the record acceptance stage, each record is checked to ensure that all necessary components were completed during the interview. This involves checking that there is demographic data for each household member and that there is labour force data for those who should have it based on their final response code, age, household membership, etc. Missing and inconsistent values of age and household membership are imputed at this stage by either carry-forward imputation or imputation by deduction.

Demographic editing involves detailed editing of the demographic information and is the final stage of editing at the household level. In this stage, all of the demographic data for all individuals in the household are edited at both the individual and within-family levels. A series of validity and consistency edits check for consistency of responses across questions for each individual and between household members. Both automated and manual corrections may be made at this stage.

The next stage of editing is the LFI item acceptance stage. In this stage, each record is run through a pre-edit process to check the validity of the labour force data received. The flow of the questions is checked to determine whether the responses for the labour force data follow a single, consistent and correct path. This process also checks the range and validity of the responses on the path.

The last stage in Phase I editing is industry and occupation coding. Records requiring coding are identified and coded using either an automated system or a manual system when the automated system cannot assign a complete code. The codes are validated and checked for consistency. Industry is coded to the North American Industry Classification System (NAICS) standard and occupation is coded to the National Occupation Classification – Statistics (NOC-S) standard. NAICS 2012 and NOC 2011 are currently used in the LFS processing system.

## 5.2 Phase II editing

Phase II editing includes resolution of 'Don't know' and 'Refusal' responses and detailed consistency editing. During this phase of editing, each record is checked to determine if it contains any entries of 'Don't know' or 'Refusal'. These responses are considered item-level nonresponse. All item-level nonresponse is identified and treated with imputation by deduction or carry-forward imputation where possible. Consistency edits are applied to ensure that each record is internally consistent. If this process does not succeed then the missing or inconsistent items are flagged for hot-deck imputation.

## 5.3 Hot-deck imputation

In hot-deck imputation for the LFS, the missing values of a recipient are replaced by the corresponding values of a randomly selected donor within the same imputation class. Imputation classes are defined based on variables available for both recipients and potential donors. Two separate sets of imputation classes are formed, one set of classes for item nonresponse, and another set for person and household nonresponse and item nonresponse which cannot be resolved through item imputation.

In January 2005, a longitudinal hot-deck imputation strategy was implemented based on research by Bocci and Beaumont (2004). This strategy is primarily used to treat person and household nonresponse. The strategy uses the previous month's values (possibly imputed) of some variables together with some socio-demographic variables from the current month as matching variables to form the imputation classes for both donors and recipients. Recently, the effectiveness of these matching variables for the treatment of person and household nonresponse was reviewed by White and Benhin (2013). The study resulted in an improved set of matching variables that was implemented in January 2015.

The following subsections describe the steps of LFS hot-deck imputation in more detail. Figure 5.1, at the end of this chapter, shows a general picture of the hot-deck imputation system (HDIS) process. More detailed specifications of the HDIS strategy can be found in Lorenz (1996).

### 5.3.1 Imputation pre-processing

Before the actual hot-deck imputation of missing values can be performed, some pre-processing steps must be completed.

First, responses are identified and each response record is assigned a preliminary imputation type. The data extracted from the head office processing system (HOPS) are divided into response and nonresponse files. The nonresponse records will be accounted for with a nonresponse weighting adjustment, which is discussed in Chapter 6, and the response records will undergo hot-deck imputation. The records in the 'response' file at this stage did not necessarily respond in the current month. If the person had responded in the previous month, then it is defined as a response in this step. All response records are initially divided into three groups: A, B and C. Group A contains potential donors. These are all persons for whom the reported data contain no missing values and are internally consistent. Group B is formed by all persons who have no missing values and are internally consistent after the first phase of editing, but do not belong in Group A because they had one or more items imputed during editing. The remaining persons form Group C and require imputation.

The second pre-processing step derives imputation matching variables. Some variables are not initially in a form which can be used directly in imputation. For example, the two occupation group variables OCC4 and OCC10 are derived from the NOC variable. Also in this step, earnings data are converted to an hourly basis by dividing total earnings for the time period by hours worked. This ensures that there is a uniform measure of earnings and that the value imputed for earnings is consistent with the value for the number of hours reported by a recipient.

The third pre-processing step is the identification of outlier earnings and the finalization of Groups A, B, and C. Earnings values that are either extremely high or low are deemed suspicious, and so they are set to missing and are imputed. Individuals who reported earnings that are very high or very low without being extreme keep those earnings values, but are excluded from being potential donors by being assigned to Group B. Outlier detection classes are formed by crossing the variables province-sex-age group, and occupation group. Different threshold values based on the quartile method are set in each class. The quartile method for outlier detection is described in *Survey Methods and Practice* published by Statistics Canada (2003).

After outlier detection, records in Group A form the potential donor pool. Records in Group C are the recipients and will be imputed by hot-deck imputation. The records in Group B do not need to be imputed and are also not eligible as donors.

The last step of pre-processing is to assign a temporary path (TPATH) to each record, where possible. This variable TPATH will be used as an important matching variable in the imputation for item nonresponse. The use of TPATH will be explained in detail in the next section.

### 5.3.2 Imputation for item nonresponse

Once all of the pre-processing steps have been completed, missing values can be imputed. Random hot-deck imputation within classes is used to fill-in missing values. This procedure is applied in such a way that the recipient data satisfy consistency edit rules and validity edit rules after imputation. For example, variables requiring non-blank values for a given recipient must be imputed using non-blank values. In a given imputation class, each recipient is imputed by selecting a series of donors using simple random sampling without replacement until a donor that satisfies all the edit rules is found. Once a suitable donor has been found, all of the recipient's missing items are imputed with data from that donor.

The initial imputation classes are formed by crossing the following eighteen categorical variables:

1. TPATH (12 categories)
2. LMLFS3 (3 categories)
3. COW (3 categories)
4. OCC4 (4 categories)
5. PROV (10 categories)
6. AGEGP3 (3 categories)
7. ABQ1 (2 categories)
8. IMM (3 categories)
9. LMLFS7 (7 categories)
10. LMINDG (20 categories)
11. MULTJOB (2 categories)
12. AGEGP1 (5 categories)
13. SEX (2 categories)
14. OCC10 (10 categories)
15. AGEGP2 (8 categories)
16. STUD (2 categories)
17. EDUC (2 categories)
18. DWELRENT (2 categories).

The order of these variables reflects their importance in explaining the labour force variables as determined by the empirical studies in White and Benhin (2013). A detailed description of the values of the categories for each variable is given in Appendix F.

Note that the variables LMLFS3, LMLFS7 and LMINDG refer to values from the last month.

The variable TPATH has an important role in the imputation system. The first seven possible values of TPATH correspond to the seven possible values of the labour force status variable, LFSSTAT. Each donor is assigned a value for LFSSTAT based on reported data. For the recipients, the value of LFSSTAT may not be known; however, there may be enough information to exclude one or more of the seven possible values. The variable TPATH is used to ensure that only valid values for LFSSTAT are imputed to recipients by replicating each donor by its number of valid TPATH values and assigning only one value of TPATH to each recipient. At the end of the imputation step, the replicated donors are removed. For example, assume that a donor has LFSSTAT = 2. This donor then has

three valid TPATH values: 2, 8, and 10 (see Appendix F for the definition of all possible TPATH values). The donor is therefore replicated three times with each replicate given one of the three valid TPATH values. When imputation classes are formed, each of the donor replicates will be in a separate imputation class.

Imputation is performed in each class that contains enough donors to pass the following two constraints:

i. The number of donors must be larger than the number of recipients of that class;

ii. Each class must contain at least three donors.

If either of these constraints is not satisfied, then the least important variable (DWELRENT) is removed from the list of imputation class variables and the imputation process is attempted again for the remaining recipients. If after this second pass of imputation there are still some recipients that have not been imputed due to classes that do not satisfy the above two constraints, then a third pass of imputation is performed by removing the second-least important variable (EDUC). This process of removing one variable followed by imputation continues until all recipients have been imputed or until only the first five variables – TPATH, LMLFS3, COW, OCC4 and PROV – remain. Any recipients not yet imputed at that point are sent for whole record imputation, in which all labour force variables of the recipient, including those that were reported, are replaced by those of a randomly selected donor using a different set of matching variables – see Section 5.3.3.

In a given imputation class satisfying the above two constraints, each recipient is imputed by first selecting a donor such that the validity edit rules are satisfied. If no such donor can be found then the record is sent for whole record imputation as described in Section 5.3.3. If a suitable donor can be found (*i.e.*, one that satisfies the validity edit rules after imputing the missing values of the recipient), the missing values of the recipient are replaced by the corresponding values from the donor and consistency edit rules are checked. If all edit rules are satisfied then the imputation process for this recipient is completed; otherwise, a second suitable donor (*i.e.*, satisfying the validity edits) is attempted and the consistency edit rules are checked again. If all edit rules are satisfied after this second attempt then the imputation process for this recipient is completed; otherwise, the entire record will be imputed using the values of the last attempted donor. This imputation for the entire record is slightly different than the imputation process described below in that it uses a different and more precise set of matching variables.

### 5.3.3 Imputation for person nonresponse

Person or household nonresponse where previous month data is available and item nonresponse that could not be treated with item nonresponse imputation is treated by whole record (longitudinal) imputation. In this strategy, data from the previous month (possibly imputed) for some variables and data from the current month for other variables are used to form the imputation classes. This strategy is also used for person and household nonresponse where there is no response in the previous month but there was a response in the past. Donors and recipients for the whole record imputation are both person-records, even when dealing with entire household nonresponse. For households where this imputation is not possible, a nonresponse weight adjustment is performed instead.

The variables currently crossed to form initial imputation classes for whole record imputation are given below in order of importance:

1. PROV (10 categories)
2. LMLFS3 (3 categories)
3. AGEGP1 (5 categories)
4. SEX (2 categories)
5. LMINDG (20 categories)
6. LMLFS7 (7 categories)
7. EIER (56 categories)
8. EDUC (2 categories)
9. ABQ1 (2 categories)

The categories for these variables are detailed in Appendix F.

As with item nonresponse imputation, the same two imputation constraints apply to the imputation classes:

i. The number of donors must be larger than the number of recipients of that class;

ii. Each class must contain at least three donors.

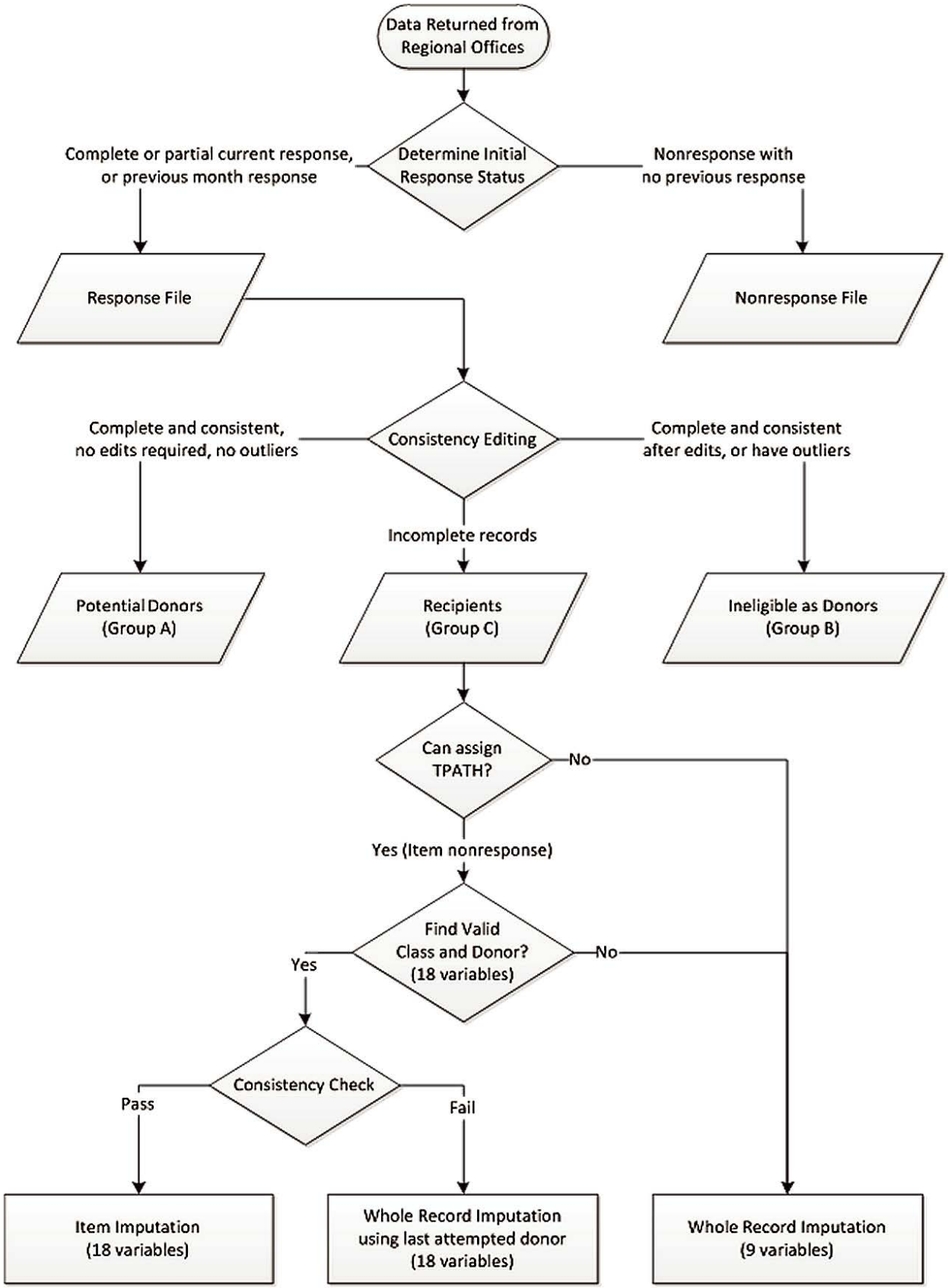
If one of these constraints is not satisfied, then classes are collapsed by removing the least important variable. The process of removing one variable and reforming the imputation classes continues until all recipients are imputed.

Recipient records are imputed using donor values from the current month, even though imputation classes are based on values from both the current month and the previous month. Also, validity and consistency edit checks are not needed when whole record imputation is performed because the donors must already satisfy all validity and consistency rules.

## 5.4 Post-imputation process

The post-imputation process includes eliminating the replicates of donors that were created during the derivation of TPATH, setting all of the group level output flags to indicate that imputation has taken place, and post-processing the earnings data by calculating both hourly and weekly earnings for all employees based on either reported or imputed hours and wages. The labour force status and some other variables are also derived.

**Figure 5.1 – Simplified Flowchart of the LFS Hot-Deck Imputation System**



# Chapter 6 Weighting and estimation

## 6.0 Introduction

Estimation is the survey process by which estimates of unknown population parameters are produced using data from a sample, possibly in combination with auxiliary information from other sources. Examples of population parameters of interest include population totals, means and ratios, as well as their averages over a number of survey months.

Labour Force Survey (LFS) estimates are produced using weights attached to each person for which LFS data is available. This chapter describes the steps involved in deriving final weights for estimation. Section 6.1 describes the calculation of design weights that reflect the sample design described in Chapter 2. Section 6.2 describes how the design weights are adjusted for nonresponding households to become what are called the subweights. Section 6.3 describes the composite calibration that is applied to the subweights to ensure consistency with external estimates of population, account for undercoverage and improve the efficiency of the estimates. This section also describes the integrated method of weighting, which ensures a common final weight for every person within a household. Finally, Section 6.4 describes how the weights are used to calculate some of the main population parameters estimated by the LFS.

## 6.1 Design weight

The design weight of a person *l* is equal to the inverse of his or her probability of being selected in the sample, $\pi_l^D$. This can be denoted by $w_l^D = 1/\pi_l^D$. The design weight is often interpreted as the number of units in the target population that the sampled unit represents. Since every person of a selected household is included in the sample, computing the selection probability of a given person is equivalent to computing the probability that the person's household is selected.

### 6.1.1 Basic weight

As described in Section 2.6.2, the overall selection probability of household *k* in PSU *j* in rotation group *i* of stratum *h* is $\pi_{hijk}^* = 1/ISR_h^*$, for all households in stratum *h*. Recall that $ISR_h^*$ is the rounded inverse sampling ratio for stratum *h*, as established during the allocation of the sample.

In all provinces except Prince Edward Island (PEI), the LFS uses a two-stage sampling design to select households. As such, the derivation of the basic weights is different for PEI than for the rest of the provinces; however, because the dwellings are selected systematically according to the stratum ISR, the selection probability in PEI is $\pi_{hijk}^* = 1/ISR_h^*$ as in the other provinces.

Since the LFS data is collected for every eligible person within a selected household, the basic selection probability of a person *l* in stratum *h* of any province is $\pi_{hijkl}^B = \pi_{hijk}^* = 1/ISR_h^*$ and his or her basic weight is

$$w_{hl}^B \equiv w_{hijkl}^B = 1/\pi_{hijkl}^B = ISR_h^*$$

This sampling design is self-weighting within strata because it has a constant basic weight within each stratum.

The design weights would be equal to the basic weights if the sampling design and the population remained unchanged. However, because the primary sampling units (PSUs) experience growth over time and the systematic sampling rate is fixed, this would lead to an ever-increasing sample size. To avoid this, the sample size is controlled through the sampling procedures described in section 3.3.2: PSUs can be sub-sampled using the PSU sub-sampling method or the sub-clustering method; the stratum can be redesigned based on updated information. These methods change the basic selection probability of households (and people). It is thus necessary to adjust the basic weights to create cluster specific weights to compensate for these sampling procedures.

### 6.1.2 Cluster weight

Cluster weights are used for strata with a two-stage design, *i.e.*, the strata for all provinces except PEI. A cluster corresponds to a PSU in these strata. In population centres, construction can cause the number of dwellings in some clusters to grow substantially over time. Interviewers are assigned clusters, and if significant growth occurs in one or more of their clusters, their workload would also grow substantially. This could affect the quality of the interviewer's work and his or her ability to complete the assignment. When the number of dwellings in a cluster increases to more than double the initial level, without becoming too extreme, the cluster may be randomly sub-sampled using either the cluster / mechanical sub-sampling or sub-clustering method. These methods of sub-sampling modify the selection probabilities of households. As a result, the basic weight $w_{hl}^{B}$ is modified by a cluster adjustment factor $a_{hl}^{P}$ to give the cluster weight

$$w_{hl}^{P} = w_{hl}^{B} a_{hl}^{P}.$$

Unfortunately, the self-weighting property is lost when either of these methods is used. Additional details of these methods can be found in Kennedy (1998). When growth is extreme, sub-sampling may not be practical, and the stratum is updated as described in below.

#### Cluster sub-sampling

This method is the simplest and most common of all subsampling methods. The sampling rate is modified to reduce the number of households selected in the cluster. If the cluster was originally sampled at a rate of $1/ISR_{hij}^{*}$ and subsampling leads to a sampling rate of $1/ISR_{hij}^{**}$, then the cluster adjustment factor is $a_{hl}^{P} = ISR_{hij}^{**} / ISR_{hij}^{*}$. The basic weights of interviewed households are multiplied by this factor. In order to use this method, the growth has to be sufficient to warrant a factor of at least two. Due to outlier problems encountered by special surveys that use the LFS frame, the maximum value of the cluster adjustment factor is three.

#### Sub-clustering

When a cluster more than triples in size and street patterns are well defined, the growth cluster is divided into 4 or more sub-clusters. A sample of 2 of these smaller sub-clusters is taken and then a sample of households is selected within each selected sub-cluster. This procedure is equivalent to adding another stage of sampling within growth clusters. It does not change the selection probability of clusters, but it does change the selection probability of households within growth clusters. The cluster sub-weight represents this selection process.

#### Stratum updates

When growth is so extreme that the sub-sampling processes described above are insufficient, then a stratum update is required, as described in Section 3.3.2. Updated counts of dwellings for all clusters in the stratum are required and new clusters are formed by sub-clustering existing clusters in the frame based on the new counts. An update to the stratum sample is implemented, based on Keyfitz (1951), as modified by Drew, Choudhry, and Gray (1978), retaining as many of the originally selected PSUs as possible. The new sample is phased-in over six months. An interim weighting factor is applied to all PSUs in the stratum until completion of the phase-in. This weighting factor adjusts for the new knowledge derived from the latest count of dwellings that is not otherwise reflected in the active sample.

### 6.1.3 Stabilization weight

The final stage of sampling is conducted using systematic sampling at a fixed rate. As the same sampling rate is used consistently over time, growth in the population, and hence in the number of households, would lead to an ever-increasing sample size and escalating survey costs if sample stabilization were not carried out.

Sample stabilization consists of randomly sub-selecting households from the sample in order to maintain the sample size at its planned level. This random selection procedure is performed using systematic sampling within each stabilization area and independently between stabilization areas. A stabilization area is defined as containing all households belonging to the same Employment Insurance Economic Region (EIER) and the same rotation group.

Sample stabilization modifies the selection probability of households. As a result, the cluster weight $w_{hl}^{P} = w_{hl}^{B} a_{hl}^{P}$ is modified by a stabilization adjustment factor $a_{hl}^{S}$ to give the stabilization weight $w_{hl}^{S} = w_{hl}^{B} a_{hl}^{P} a_{hl}^{S}$. By definition, the design weight of a person *l* in stratum *h*, $w_{hl}^{D}$, is equal to its stabilization weight $w_{hl}^{S}$, *i.e.*,

$$w_{hl}^{D} \equiv w_{hl}^{S} = w_{hl}^{B} a_{hl}^{P} a_{hl}^{S}.$$

**Calculating the stabilization adjustment**

The stabilization adjustment factor $a_{hl}^{S}$ is computed separately within sub-areas. A sub-area is defined as all strata within a stabilization area that have a common sampling fraction. Stabilization weighting departs slightly from the principle of weighting by the inverse of the selection probability since it is performed within sub-areas and not within stabilization areas. Such a weighting procedure is often called poststratification, with the poststrata being the sub-areas in this case.

To give a simplified example, suppose that there is a stabilization area in which all households have a basic selection probability of 1 in 200 at the time of design and a common cluster adjustment factor of 1. In this simplified example, the stabilization area is thus not partitioned into sub-areas. If the stabilization area has a planned sample size of 300 households at the time of design, and if the sampling rates used in fact yield 350 households, then 50 households must be dropped randomly from the stabilization area. This changes the selection probability of households from 1 in 200 to 3 in 700 (*i.e.*, 1/200 times 300/350). The basic weight of 200 is thus multiplied by the factor 350/300 to yield the stabilization weight 700/3=233.333333.

Households that have one of the following two characteristics are excluded from sample stabilization and stabilization weighting:

- Households belonging to a cluster that has been subsampled using cluster sub-sampling or sub-clustering as described in Section 6.1.2;
- Households living in a recently-built dwelling, which has been added to the cluster list and was thus not eligible to be dropped (interviewer selected dwelling).

Since such households do not get a chance to be dropped from the sample, they are excluded from stabilization weighting as well.

## 6.2 Subweight

While an attempt is made to interview all households in the selected sample *s*, refusals and other factors make it impossible to contact some households. Part of this household nonresponse is first treated by using a longitudinal imputation method (see Section 5.3.3). Then, the remaining nonrespondent households are treated by removing them from the file and adjusting the design weights of responding households, including those that have been imputed, by a nonresponse adjustment factor. The basic principle consists of determining an appropriate model for the unknown response probabilities and then computing the nonresponse adjustment factors as the inverse of the estimated response probabilities.

In the LFS, the nonresponse model used is the uniform nonresponse model within classes. With this model, all households within a given nonresponse class *c* are assumed to have the same response probability $p_c$. The estimated response probability $\hat{p}_c$ is simply the design-weighted response rate of households within class *c*. The nonresponse adjustment factor for a person *l* belonging to a responding household in class *c* is $a_{cl}^{NA} = 1/\hat{p}_c$ and the nonresponse adjusted weight, or the subweight, is

$$w_{cl}^{NA} = w_{cl}^{B} a_{cl}^{P} a_{cl}^{S} a_{cl}^{NA} = w_{cl}^{D} a_{cl}^{NA}.$$

Every person within a given responding household has the same nonresponse adjustment factor and thus the same subweight.

### 6.2.1 Nonresponse classes

The key to reducing nonresponse bias is to determine nonresponse classes that explain the unknown nonresponse mechanism well and that are constructed in such a way that the assumption of constant response probability within classes is reasonable. From an efficiency point of view, it is also desirable that nonresponse classes be as homogeneous as possible with respect to the main variables of interest, that is, classes should be formed in such a way that the respondents within a given class are similar to nonrespondents in terms of the main variables of interest. As a result, variables used to construct classes should be explanatory for the nonresponse mechanism and also for the main variables of interest.

In the LFS, every aboriginal or high-income stratum forms a separate nonresponse class. The remaining classes are defined by crossing the variables PROVINCE, EIER, TYPE and ROTATION (excluding households belonging to an Aboriginal or high-income class). The variable TYPE has four categories and indicates the type of stratum to which a household belongs: Remote, Rural, Urban non-Census Metropolitan Area (CMA) (including PEI one-stage strata) and Urban CMA. The variable ROTATION corresponds to the six rotation groups. Note that the nonresponse classes do not overlap and, collectively, they cover the entire population. Collapsing of classes is performed when a nonresponse adjustment factor is greater than two in a given class. This is done by removing the last class variable (ROTATION) and recalculating the nonresponse adjustment factors among the redefined classes (PROVINCE by EIER by TYPE). The problematic class then gets the new adjustment factor, as well as all other classes (*i.e.* rotation groups) within the same PROVINCE, EIER and TYPE. The reason for collapsing nonresponse classes is to avoid large nonresponse adjustment factors since they tend to increase the variability of the estimates.

## 6.3 Final weight

The last step of the weighting process is to derive the final weights, which are used to obtain official estimates. Composite calibration and the integrated method of weighting are used to produce the final weights. The integrated method of weighting is used to ensure a common final weight for every person in the household.

### 6.3.1 Composite calibration

Calibration is used for the following three reasons: to ensure consistency with Census projected estimates and with all surveys using these Census estimates; to account for undercoverage; and to improve the efficiency of the estimates. To account for undercoverage and improve the efficiency of the estimates, auxiliary variables used in calibration must be correlated with the main variables of interest. One way to achieve this goal is to choose auxiliary variables by modelling the variables of interest. For example, an appropriate model can show that being employed or unemployed is related to the age and sex of a person.

The LFS uses composite calibration (or regression composite estimation) to produce the final weights. Composite calibration is essentially the same as calibration, except that some control totals are estimates from the previous month's survey data and the auxiliary variables associated with these control totals are imputed for some units.

Composite calibration can lead to substantial improvement in the efficiency of the estimates if there is a strong month-to-month correlation in the information collected. Such improvement is due to the overlapping nature of the LFS sample. On the one hand, gains in efficiency are obtained because composite calibration uses information obtained in the previous month from the exit rotation group. On the other hand, it also has a reduction in efficiency due to missing values in the birth rotation group. Overall, it was found empirically that composite calibration is beneficial in the LFS.

Like calibration, composite calibration is a technique that finds weights $w_l^{CC}$, for all people in the subset of all people from the sample, *s*, who belong to a responding or imputed household, $l \in s_r$, as close as possible to the subweights $w_l^{NA}$, subject to some constraints. More formally, composite calibration weights, $w_l^{CC}$, are obtained in the LFS by minimizing the distance function

$$\sum_{l \in s_r} \frac{\left(w_l^{CC} - w_l^{NA}\right)^2}{w_l^{NA}} \qquad (6.1)$$

subject to two sets of constraints: calibration constraints, and composite calibration constraints.

The first set of constraints, the calibration constraints, require that estimates based on the weights, $w_l^{CC}$, for a vector of auxiliary variables $\mathbf{x}$, $\hat{\mathbf{X}}^{CC} = \sum_{l \in s_r} w_l^{CC} \mathbf{x}_l$, are equal to the vector of known population totals, $\mathbf{X} = \sum_{l \in P} \mathbf{x}_l$. In other words, the calibration constraints can be given by $\sum_{l \in s_r} w_l^{CC} \mathbf{x}_l = \mathbf{X}$. In the LFS, these known population totals, often called control totals, are Census estimates projected to the current month for the number of people aged 15 and over in Economic Regions (ERs) and CMAs/Census Agglomerations (CAs), and for the number of people in 24 age-sex groups by province. Additional control totals are used to ensure that the estimated number of people aged 15 and over is the same for each rotation group. To perform calibration, the vector $\mathbf{x}$ must be known for every person $l \in s_r$. In the case of the LFS, this means that the age-sex group, ER, and CMA/CA of each person $l \in s_r$ must be known.

The second set of constraints, the composite calibration constraints, involve control totals that are estimates from the previous month's survey data, and auxiliary variables associated with these estimated control totals. The auxiliary variables may not be known for all people $l \in s_r$ and are thus imputed for some. These control totals and auxiliary variables are called composite control totals and composite auxiliary variables respectively. There are 28 composite auxiliary variables for each province and they are all defined with respect to the previous month's survey data (see Appendix G for a complete list).

### Imputation of auxiliary control variables

If the vector of composite auxiliary variables for unit *l*, denoted by $\mathbf{z}_{t-1,l}$, is defined for the previous month (month $t-1$), the corresponding vector of estimated control totals, denoted by $\hat{\mathbf{Z}}$, must also be computed using the previous month's data. The vector of composite auxiliary variables $\mathbf{z}_{t-1}$ is not observed for people in the birth rotation group since they were not interviewed in the previous month. Imputation is used to fill in missing values for these units using a combination of two imputation methods.

In the first method, mean imputation is used to obtain the modified vector:

$$z_{\bullet l}^{(1)} = \begin{cases} \mathbf{z}_{t-1,l} & \text{if } l \in s_r - s_r^b \\ \hat{\mathbf{Z}} / N_{15+} & \text{if } l \in s_r^b \end{cases},$$

where $s_r^b$ is the subset of people $l \in s_r$ who belong to the birth rotation group and $N_{15+}$ is the provincial number of people aged 15 and over. In a previous empirical study, it was found that this imputation method was efficient for estimating population parameters defined at the current month *t*.

In the second imputation method, the modified vector $\mathbf{z}_{\bullet l}^{(2)}$ is defined as:

$$\mathbf{z}_{\bullet l}^{(2)} = \begin{cases} \mathbf{z}_{t-1,l} + \left(\delta_l^{-1} - 1\right)\left(\mathbf{z}_{t-1,l} - \mathbf{z}_{t,l}\right) & \text{if } l \in s_r - s_r^b \\ \mathbf{z}_{t,l} & \text{if } l \in s_r^b \end{cases},$$

where $\mathbf{z}_{t,l}$ is the vector $\mathbf{z}_{t-1,l}$ defined at the current month *t* and $\delta_l$ is the probability that $l \in s_r - s_r^b$ given that $l \in s_r$. In the LFS, $\delta_l = 5/6$, for $l \in s_r$, and is replaced in the previous equation by the estimate $\hat{\delta}_l = \sum_{l \in s_r - s_r^b} w_l^{NA} \Big/ \sum_{l \in s_r} w_l^{NA}$. Essentially, the idea is to perform carry-backward imputation (imputation by current month's values to fill in previous month's values) to impute $\mathbf{z}_{t-1}$ for the birth rotation group since it is known that there is a strong month-to-month correlation for the composite auxiliary variables. However, the values of $\mathbf{z}_{t-1}$ in the non-birth rotation groups are modified due to the fact that carry-backward imputation eliminates change for people in the birth rotation group. The correction in the non-birth rotation group is determined so as to preserve the property of asymptotic unbiasedness of the estimates. In a previous empirical study, it was found that this imputation method (which determines $\mathbf{z}_{\bullet l}^{(2)}$) was efficient for estimating population parameters defined as differences between two successive months.

As stated, neither $\mathbf{z}_{\bullet l}^{(1)}$ nor $\mathbf{z}_{\bullet l}^{(2)}$, is actually used in the survey. Instead, a combination of the two methods is used. The composite auxiliary variables are defined as

$$\mathbf{z}_{\bullet l} = (1-\alpha)\mathbf{z}_{\bullet l}^{(1)} + \alpha \mathbf{z}_{\bullet l}^{(2)},$$

where $\alpha$ is a tuning constant that equals 2/3. This leads to a compromise between the two imputation methods. A study on the choice of $\alpha$ can be found in Chen and Liu (2002). Alternative imputation methods have also been studied in Bocci and Beaumont (2005) using the idea of calibrated imputation.

The LFS composite calibration weights $w_l^{CC}$ are therefore obtained by minimizing the distance function given by Equation (6.1), subject to both sets of constraints

$$\sum_{l \in s_r} w_l^{CC} \begin{pmatrix} \mathbf{x}_l \\ \mathbf{z}_{\bullet l} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \hat{\mathbf{Z}} \end{pmatrix}$$

The minimization leads to the composite calibration weights $w_l^{CC} = w_l^{NA} g_l^{CC}$ where the composite calibration adjustment factor $g_l^{CC}$ is given by

$$g_l^{CC} = \left(\mathbf{x}_l', \mathbf{z}_{\bullet l}'\right)\left(\sum_{l \in s_r} w_l^{NA}\left(\mathbf{x}_l', \mathbf{z}_{\bullet l}'\right)'\left(\mathbf{x}_l', \mathbf{z}_{\bullet l}'\right)\right)^{-1}\left(\mathbf{X}', \hat{\mathbf{Z}}'\right)'.$$

Additional details about LFS composite calibration can be found in Singh, Kennedy and Wu (2001), Fuller and Rao (2001) and Gambino, Kennedy and Singh (2001). Gambino, Kennedy and Singh (2001) also discuss issues related to missing and out-of-scope people at the previous month in the non-birth rotation groups. Missing values are imputed using random hot-deck imputation and $\mathbf{z}_{\bullet l} = \mathbf{0}$ is assigned to out-of-scope people at the previous month. The idea is to determine $\mathbf{z}_{\bullet l}$ so that $\sum_{l \in s_r} w_l^{NA} \mathbf{z}_{\bullet l}$ remains, like $\hat{\mathbf{Z}}$, an estimate of the unknown vector of

control totals $\mathbf{Z}$, which is defined for the previous month. Missing values and out-of-scope people at the current month are dealt with in the usual way.

### 6.3.2 Integrated method of weighting

Since some auxiliary variables and all composite auxiliary variables are defined at the person level, the composite calibration weights $w_l^{CC}$ are not constant within a household, unlike the subweights $w_l^{NA}$. This does not pose a problem as long as the interest is in estimating person-related population parameters, such as the total number of people employed in the population. However, in the LFS, there is also sometimes interest in estimating household-related population parameters. For example, there may be interest in estimating the total number of households having a certain characteristic, such as having at least one member employed. There is more than one weighting alternative for such population parameters.

In order to avoid producing two sets of final weights, the integrated method of weighting was introduced in the LFS to obtain a unique set of final weights that can be used for both person-related and household-related population parameters; see Lemaître and Dufour (1987). With this method, the final composite calibration weight is constant for all the people within a household. This is achieved by replacing $\mathbf{x}_l$ and $\mathbf{z}_{\bullet l}$ for a given person l by the average of $\mathbf{x}$ and $\mathbf{z}_{\bullet}$ over all members of his or her household and then computing the composite calibration weights as in Section 6.3.1. This ensures a common final weight for all people within the same household. This additional constraint on the final weights is expected to reduce the efficiency of the estimates. However, Pandey, Alavi and Beaumont (2003) have found empirically that the reduction in efficiency is small in the context of the LFS.

### 6.3.3 Treatment of negative weights and rounding

Sometimes calibration results in negative weights. In this situation, composite calibration is performed again on the post-calibration weights, with the negative weights reset to their subweights. If after this second round of composite calibration there are still negative weights, then these negative weights are set equal to 1 and it is accepted that the composite calibration constraint will not be perfectly satisfied. This rarely occurs. After both rounds of composite calibration the weight is rounded to the nearest integer, producing the final weight.

## 6.4 Estimation

Once the final weights have been calculated, they are used to estimate several types of population parameters, including the following examples of totals, rates and moving averages.

Each month, the LFS calculates the number of employed people in the population. If $y_l$ is a binary variable indicating whether a given person *l* of the population is employed $\left(y_l = 1\right)$ or not $\left(y_l = 0\right)$, the population total *Y* represents the number of employed people in the population *P*. The population total is calculated as

$$Y = \sum\nolimits_{l \in P} y_l$$

Using the final weights, this population total can be estimated by

$$\hat{Y}^{CC} = \sum\nolimits_{l \in s_r} w_l^{CC} y_l$$

where $S_r$ is the subset of all the people from *s* who belong to a responding or imputed household and $w_l^{CC}$ is the composite calibration weight, or final weight, attached to person *l*.

The LFS also calculates the unemployment rate each month. If $y_{1l}$ is a binary variable indicating whether a given person *l* of the population is unemployed $\left(y_{1l} = 1\right)$ or not $\left(y_{1l} = 0\right)$ and $y_{2l}$ is a binary variable indicating

whether person *l* is in the labour force $\left(y_{2l}=1\right)$ or not $\left(y_{2l}=0\right)$, then the population rate $r_{y_1,y_2}$ represents the unemployment rate in the population.

$$r_{y_1,y_2}=\frac{\sum_{l\in P} y_{1l}}{\sum_{l\in P} y_{2l}}.$$

It can be estimated using the final weights $w_l^{CC}$ by

$$\hat{r}_{y_1,y_2}^{CC}=\frac{\sum_{l\in s_r} w_l^{CC} y_{1l}}{\sum_{l\in s_r} w_l^{CC} y_{2l}}.$$

As well, every month, the LFS produces three-month moving average estimates of the unemployment rates for each EIER using data from the three most recent months. If the T-month moving average of a total $Y$ at time *t* is

$$\theta_t^Y=\sum_{q=0}^{T-1}\frac{Y_{t-q}}{T}$$

and it is estimated using the final weights by

$$\hat{\theta}_t^Y=\sum_{q=0}^{T-1}\frac{\hat{Y}_{t-q}^{CC}}{T}$$

then the estimated three-month moving average for the unemployment rate can be calculated as

$$\hat{r}_{\theta_t^{Y_1},\theta_t^{Y_2}}=\frac{\hat{\theta}_t^{Y_1}}{\hat{\theta}_t^{Y_2}}=\sum_{q=0}^{2}\frac{\hat{Y}_{1,t-q}^{CC}}{3}\Bigg/\sum_{q=0}^{2}\frac{\hat{Y}_{2,t-q}^{CC}}{3}$$
$$=\sum_{q=0}^{2}\hat{Y}_{1,t-q}^{CC}\Bigg/\sum_{q=0}^{2}\hat{Y}_{2,t-q}^{CC}$$

Moving average estimates are used because they are more stable than monthly estimates; however, their interpretation is different since they estimate a different population parameter.

# Chapter 7 Variance estimation

## 7.0 Introduction

In a survey based on a probability sample such as the Labour Force Survey (LFS), statistical inferences need to account for the sampling error. The variance measures the precision of an estimator. Because of the complexity of the estimation method and sample design, an explicit form of the variance estimator is not readily available for the LFS. The survey therefore uses a resampling method for variance estimation.

With the 2015 redesign, a major change to the LFS variance estimation methodology was introduced. Previously, variance estimation was based on a resampling method called the jackknife. A variance estimation system custom-built for the LFS used the jackknife method to produce variance estimates of totals, rates or proportions, changes, and moving averages. As of January 2015, variance estimation is based on a resampling method called the bootstrap. Each month, 1,000 sets of LFS bootstrap weights are generated, and these bootstrap weights can be used with various standard software packages to produce variance estimates. The variance estimates obtained using the new methodology are similar in value to those obtained using the old methodology. The main advantage of the new methodology is that once bootstrap weights are generated, they can be used to produce variance estimates for a much wider variety of analyses than the old system.

This chapter will describe how variance is estimated for the LFS. Section 7.1 presents the particular bootstrap method that is implemented, the Rao-Wu bootstrap. Sections 7.2 and 7.3 describe how the LFS bootstrap samples and bootstrap weights are generated. Section 7.4 discusses how the bootstrap weights are used to compute variance estimates.

## 7.1 The Rao-Wu bootstrap

The LFS uses the Rao-Wu bootstrap, as proposed in Rao and Wu (1988) and Rao, Wu and Yue (1992). The method was proposed for stratified multistage designs where the primary sampling units (PSUs) are selected using probability proportional to size with replacement (PPSWR) sampling. For the LFS, the PSUs are actually selected using PPS without replacement (PPSWOR). Särndal, Swensson and Wretman (1992, p. 154), states that the variance estimator for multistage sampling with PSUs selected without replacement can be approximated by the variance estimator for multistage sampling with PSUs selected with replacement, and that the approximation is conservative if the selection of PSUs without replacement is more efficient than the selection of PSUs with replacement. This is the case for the LFS.

The first step in applying the Rao-Wu bootstrap is to select bootstrap samples. For each stratum $h$, $m_h$ PSUs are drawn using simple random sampling with replacement (SRSWR) from the original set of $n_h$ sampled PSUs.

For most applications of the Rao-Wu bootstrap at Statistics Canada, including the LFS, $m_h$ is set to $n_h - 1$.

This process of selecting bootstrap samples is repeated $B$ times. The number of times the $j^{th}$ PSU is selected in the bootstrap sample of the $b^{th}$ replicate, called the multiplicity of the PSU, is denoted as $m_{hj}^{(b)}$, where $b$=1,...,$B$. The multiplicities, $m_{hj}^{(b)}$, have values between 0 and $n_h - 1$ inclusive, and satisfy $\sum_{j=1}^{n_h} m_{hj}^{(b)} = n_h - 1$ for each bootstrap replicate and each stratum.

The next step is to produce $B$ sets of bootstrap weights by applying an adjustment factor to the original survey weight. They are calculated as follows:

$$w_{hjk}^{(b)} = \frac{n_h}{n_h - 1} m_{hj}^{(b)} w_{hjk}, \qquad (7.1)$$

where $w_{hjk}$ is the survey weight for unit *k* in PSU *j* and stratum *h*, and $w_{hjk}^{(b)}$ is the bootstrap weight for the *b*th replicate.

The *B* sets of bootstrap weights can be used to produce variance estimates for a variety of analyses. For an estimate, $\hat{\theta}$, of a population parameter, $\theta$, the bootstrap variance estimate is computed as follows. The estimate is calculated using each set of bootstrap weights, resulting in *B* estimates denoted as $\hat{\theta}^{*(1)},...,\hat{\theta}^{*(B)}$. For example, suppose $\hat{\theta}$ is an estimate of a total, given by $\hat{\theta} = \sum_h \sum_j \sum_k w_{hjk} y_{hjk}$, where $y_{hjk}$ is the value of a variable of interest *y* for unit *k* in PSU *j* and stratum *h*. Then the estimate for the $b^{th}$ bootstrap replicate is $\hat{\theta}^{*(b)} = \sum_h \sum_j \sum_k w_{hjk}^{(b)} y_{hjk}$. The bootstrap variance estimate is given by the variance of the *B* estimates

$$\hat{V}_{\text{BOOT}}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \hat{\theta}^{*(.)}\right)^2, \qquad (7.2)$$

where $\hat{\theta}^{*(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$.

## 7.2 LFS bootstrap samples

To obtain stable variance estimates for various types of analyses, as many bootstrap replicates as possible should be made available. A compromise has to be reached between ensuring stability, and limiting the execution time and the size of files. The LFS has opted to generate 1,000 LFS bootstrap replicates each month. This ensures the stability of the variance estimates for the key survey estimates.

As described in Section 7.1, the first step in applying the Rao-Wu bootstrap consists in drawing 1,000 bootstrap samples at the PSU level, with $n_h - 1$ PSUs selected with replacement per stratum. A two-stage sample design is used for all provinces except Prince Edward Island (PEI), and the bootstrap samples are therefore selected at the cluster level. Since a one-stage sample design is used for PEI, the bootstrap samples are selected at the dwelling level.

The remainder of this section provides details on various considerations related to the generation of the LFS bootstrap samples. This is followed by Section 7.3, which describes the generation of the 1,000 sets of LFS bootstrap weights.

### 7.2.1 Strata with one selected PSU

To estimate the variance, each stratum should contain at least two sampled PSUs. This is usually the case for the LFS and it is always the case for the one-stage strata in PEI. Most two-stage strata in the provinces contain six sampled PSUs, one for each rotation. However, for various reasons, some strata may only have one sampled PSU on the final tabulation file. This can happen by design (a few three-stage strata in the previous design, transition between redesigns), or due to survey results (out-of-scope and non-responding dwellings). The single-PSU strata are handled in one of three different ways.

First, the three-stage strata in the provinces with only one selected PSU are handled by splitting the selected PSU. The PSU is split by the rotation group or by the second sampling stage unit (SSU). For these strata, the bootstrap samples are selected at the rotation group level or at the SSU level instead of the PSU level.

Second, the single-PSU strata that occur during the redesign transition period are handled by collapsing strata; this is discussed in more detail in Section 7.2.3.

Finally, the remaining single-PSU strata are handled by temporarily splitting the PSU into two parts based on whether the household identifier is even or odd. The strategy was chosen because it is easy to implement and

requires no manual intervention. This situation happens rarely enough that the strategy used has no impact on the variance estimates at the provincial level.

### 7.2.2 Bootstrap sample coordination

The LFS produces estimates involving multiple survey months, such as estimates of change between periods and moving averages. The sample overlap and dependence that exists between months can be taken into account in the variance estimation through the coordinated bootstrap method proposed by Roberts, Kovacevic, Mantel and Phillips (2001). Their method takes the dependence into account by retaining the same bootstrap samples of PSUs from one month to the next.

In practice, the sampled PSUs in a stratum are not always the same from one month to the next, and the coordinated bootstrap needs to be adapted. A strategy is proposed in Neusy (2013) and Benhin and Mantel (2012) to adapt the coordinated bootstrap in the presence of change. There are potentially four different situations:

i. When the sampled PSUs in the stratum are the same in the current month as in the previous month, the previous month's bootstrap sample can be used for the current month without any further work.

ii. When the PSUs are not all the same but the number of sampled PSUs in the stratum remains the same for the two months, the coordinated bootstrap can be implemented by pairing each PSU in the current month with a PSU in the previous month. PSUs that are common to both months' samples are paired; new PSUs replacing retired PSUs are paired with the PSU that they are replacing; and all remaining PSUs are randomly paired. The current-month bootstrap samples for the stratum are generated by transferring the multiplicities of the previous month to the current month: each current-month PSU receives the multiplicities of the previous-month PSU with which it is paired. This results in a Rao-Wu bootstrap sample with the same multiplicities in the current month as in the previous month for the PSUs that are common to both months.

iii. When there are fewer sampled PSUs in the stratum for the current month than for the previous month, the coordinated bootstrap is adapted as follows. Each PSU in the current month is first paired with a PSU in the previous month as described in ii, leaving one or more previous-month PSUs unpaired. Each current-month PSU receives the multiplicities of the previous-month PSU with which it is paired, resulting in preliminary bootstrap samples for the current month. The sum of the multiplicities for the preliminary bootstrap samples is not necessarily $n_h - 1$ for all the bootstrap replicates. This is because the multiplicities of the unpaired previous-month PSUs are not carried forward to the current month and because $n_h$ is smaller than it was in the previous month. For the bootstrap replicates where the sum is less than $n_h - 1$, PSUs are randomly added to the bootstrap sample using SRSWR (i.e., the PSU multiplicities are increased) until the sum of the multiplicities is $n_h - 1$. Conversely, for the bootstrap replicates where the sum is greater than $n_h - 1$, PSUs are randomly dropped from the bootstrap sample (i.e., the PSU multiplicities are decreased) until the sum of the multiplicities is $n_h - 1$.

iv. When there are more sampled PSUs in the stratum for the current month than for the previous month, an extra step is required to adapt the coordinated bootstrap. The current-month PSUs are paired, as many as possible, with the previous-month PSUs as described in ii. The current month has more sampled PSUs than the previous month so not all current month PSUs can be paired. The paired current-month PSUs receive their multiplicities from the previous-month PSUs with which they are paired. The multiplicities of the unpaired current-month PSUs (new PSUs) are generated using the $\text{Binomial}\left(n_h^* - 1, 1/n_h^*\right)$ distribution, where $n_h^*$ is the number of sampled PSUs in the previous month. This ensures that the expected multiplicities of the unpaired PSUs are the same as the paired PSUs. The multiplicities for the paired and unpaired current month PSUs together form preliminary bootstrap samples for the current month. The sum of the multiplicities for the preliminary bootstrap samples is not necessarily $n_h - 1$ for all the bootstrap replicates. PSUs are randomly added or dropped from the bootstrap samples, as described in iii, until the sum of the multiplicities is $n_h - 1$ for all bootstrap samples.

The strategies for handling increases or decreases in the number of sampled PSUs described in iii and iv maintain correct cross-sectional variance estimates, and provide some coordination for variance estimates involving multiple months.

For the LFS, the coordination for two and three-stage strata in the provinces is implemented as follows. The LFS bootstrap samples are based on the PSUs present in the current month's final tabulation file, and the number of PSUs within most strata remains the same from one month to the next. This means that the coordination described in i and ii are most commonly used. However, there are sometimes differences in the number of PSUs, usually caused by a PSU with temporarily no respondents in the final tabulation file. If the number of PSUs decreases by one, then the adaptation to the coordinated bootstrap described in iii is used. If the number of PSUs decreases by more than one or if it increases, then new bootstrap samples are randomly selected using a fixed random seed that is assigned to each LFS stratum and kept until the next redesign. These fixed random seeds are used so that the same bootstrap samples are selected for a given stratum and number of PSUs.

Starting with the 2015 redesign, PEI uses a one-stage sample design, and PSUs are at the dwelling level. As described in Section 2.5.6, the PEI strata were formed based on the Census Dissemination Areas (DAs), and then assigned to one of six rotation groups. All dwellings in the same one-stage stratum belong to the same rotation group. Every six months, when a new sample of dwellings is rotated into a stratum, new bootstrap samples are also selected for that stratum. For the other five months, the bootstrap samples are coordinated using the strategies described in i, ii, iii and iv, depending on the situation.

### 7.2.3 Redesign transition period

The LFS sample was redesigned in January 2015 and the new sample was gradually phased-in from January to June 2015. Each month during the transition period, a rotation group from the old design was rotated out and replaced by a rotation group from the new design. In the end, the estimates are based on an integration of the two designs. The new LFS sample was selected independently from the old sample, so the bootstrap samples for the new design were also selected independently and separately from the bootstrap samples for the old design, i.e. without coordination.

As the old sample was gradually phased-out during the transition period, the number of sampled PSUs in the old design strata decreased each month. The bootstrap samples for the old design were coordinated during this period using the coordination strategy described in iii of Section 7.2.2. By the fifth month of the transition period, only one rotation group from the old design remained in the LFS sample, and therefore many strata were left with only one PSU. The single-PSU strata were randomly paired within the province and collapsed to form two-PSU strata. Preliminary bootstrap samples for the collapsed strata were generated using the previous month multiplicities of each PSU. Each collapsed stratum had two PSUs, so PSUs were randomly added or dropped from the bootstrap samples until the sum of the multiplicities was one for all the bootstrap samples of each collapsed stratum.

The bootstrap samples for the new sample were created by first generating bootstrap samples for the June 2015 sample, when the new sample was completely phased-in. The bootstrap samples were generated using a new set of random seeds that will be kept until the next redesign. Next, bootstrap samples for the new design were generated moving backwards from May to January 2015. While moving backwards, the number of sampled PSUs decreases each month. The same methodology used to coordinate the bootstrap samples for the old design moving forward through time was used to coordinate the bootstrap samples for the new design moving backward through time. The January 2015 sample contained only one rotation group from the new sample, and therefore many new design strata contained only one PSU. The single-PSU strata were collapsed and handled, as described previously for the old design strata in the fifth month of the transition.

## 7.3 LFS bootstrap weights

In order to properly estimate the sampling variability of an estimator, each of the weighting steps leading to the computation of the final weights should be repeated for each bootstrap replicate. Currently, only the final weighting step, composite calibration (see Section 6.3.1), is repeated for each bootstrap replicate. This was also the case for the previous variance estimation system based on the jackknife.

The following steps are performed to generate 1,000 sets of final LFS bootstrap weights for the provinces:

1. Initial bootstrap weights are generated for each household by applying Equation (7.1), using the multiplicities from the 1,000 LFS bootstrap samples and the household subweights from the current month LFS final tabulation file.
2. A separate set of composite control totals is required for each bootstrap replicate. The 1,000 sets of totals are calculated using the previous month's final bootstrap weights. To do this, first, each set of replicate weights from the previous month's bootstrap weight file is calibrated to the current month's demographic control totals. Next, for each set of weights, provincial-level estimates for the 28 labour characteristics listed in Appendix G are calculated.
3. Composite auxiliary variables that correspond to the composite control totals are derived for the current month households, as described in Section 6.3.1. The characteristics are derived using the previous month's final tabulation file for households that are common to both the current and previous month. The auxiliary variables of households missing from the previous month's final tabulation file are imputed using donor imputation in the case of nonrespondents, and using mean imputation in the case of households from the birth rotation. The donor imputation for the nonrespondents is only performed once, whereas the mean imputation is performed separately for each of the 1,000 bootstrap replicates.
4. The initial bootstrap weights generated in step 1 are calibrated to the current month's demographic control totals and to the composite control totals computed in step 2, using the composite auxiliary variables derived in step 3. The calibration is repeated for each bootstrap replicate.

Note that if negative weights are obtained in step 4, then the calibration is applied a second time to the calibrated weights, with the negative weights replaced by their value in the initial bootstrap weights file. If after this second round of calibration there are still negative weights, these negative weights are set to one and it is accepted that the control totals will not be satisfied.

Monthly LFS bootstrap weights have been generated beginning from 1998 for the ten provinces. They are now generated every month, as part of monthly production.

## 7.4 Variance estimation

The LFS bootstrap weights are used to compute variance estimates using Equation (7.2). The variance estimates can be produced using software packages, such as SAS (PROC SURVEYMEANS), Stata 9 or newer, SUDAAN and WesVar. Gagné, Roberts, and Keown (2014) and Phillips (2004) provide guidance on how to use bootstrap weights with these software packages.

In order to reduce the size of the LFS bootstrap weight files, the files contain one record per household. Like the survey weights, the bootstrap weights are the same for all household members, and so a person level bootstrap weights file can be generated by assigning the household level bootstrap weights to each member of the household.

As described in Section 7.1, the bootstrap variance estimate for an estimate, $\hat{\theta}$, is obtained by first computing the estimate with each set of bootstrap weights to obtain $\hat{\theta}^{*(1)},...,\hat{\theta}^{*(1,000)}$, and then applying (7.2).

For estimates involving multiple survey months, each of $\hat{\theta}^{*(1)},...,\hat{\theta}^{*(1,000)}$ should be computed using multiple survey months as well. For example, consider an estimate of change of the form: $\hat{\theta}_C = \hat{\theta}_2 - \hat{\theta}_1,$ where $\hat{\theta}_1$ is an estimate of $\theta_1$, the population parameter for the first month; and $\hat{\theta}_2$ is an estimate of $\theta_2$, the population parameter for the second month. The bootstrap variance estimate of $\hat{\theta}_C$ is obtained by first computing $\hat{\theta}_C^{*(b)} = \hat{\theta}_2^{*(b)} - \hat{\theta}_1^{*(b)}$ for *b*=1,...,1000, where $\hat{\theta}_1^{*(b)}$ is an estimate of $\theta_1$ based on the *b*th set of first month bootstrap weights, and $\hat{\theta}_2^{*(b)}$ is an estimate of $\theta_2$ based on the *b*th set of second month bootstrap weights. Equation (7.2)

is then applied to $\hat{\theta}_C^{*(1)},\ldots,\hat{\theta}_C^{*(1,000)}$. Because the bootstrap weights are based on coordinated bootstrap samples, this approach of handling estimates involving multiple periods will take into account the overlap and dependence that exists between months. In practice, software packages are not usually designed to deal with multiple datasets from different periods. A solution to this problem is to create an input file containing the data and bootstrap weights from all the months of interest in the same file. It may be necessary to create dummy variables to identify the different months.

# Chapter 8 Data quality

## 8.0 Introduction

The data quality evaluation refers to the process of evaluating the final product of the survey against the original objectives of the survey. In particular, the evaluations are in terms of data accuracy and reliability. Such information allows users to make more informed interpretation and use of the survey results. Users must be provided with information allowing them to assess the degree to which data limitations restrict the use of the data. Data quality evaluations are also of benefit to the statistical agency. When data limitations can be traced to specific steps in the survey process, such evaluations can be used to improve the quality of subsequent occasions of the survey and of other similar surveys.

The *accuracy* of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of statistical error and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (*e.g.*, sampling errors and non-sampling errors). This is the approach that will be used here.

In a sample survey, inferences are made about the target population based on the data collected from only a portion of this population. The results will probably differ from those obtainable from a complete census of this population under the same conditions. The error caused by applying conclusions to the entire population based on only a sample is called a sampling error. Some factors that contribute to sampling errors include sample size, variability of the characteristics examined, the sampling plan, and the estimation method.

Non-sampling error, as its name indicates, is not caused by the sampling process and can take place in a census or a sample survey. This type of error can occur at any step of the survey (planning, design, data collection, coding, data capture, editing, estimation, analysis, and dissemination of data) and is mainly caused by human error. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors. Non-sampling error is also associated with other types of errors, such as errors in the information sources, the methods used to obtain population projections, seasonal adjustment errors, *etc*.

To monitor and ensure the quality of its data, the LFS adopted a program to measure data quality. A range of quality indicators are regularly produced, and carefully analyzed. If there are unusual values, the LFS managers are immediately notified so they can make the necessary corrections as quickly as possible. Some indicators are merely monitored, since their role is to detect trends or long-term effects. For example, some measure the consequences of certain operational changes, while others measure the impact of minor changes to the sample design. This long-term information on data reliability can be used to make changes that are likely to improve the overall quality of the results and to help analysts and data users at Statistics Canada and elsewhere with their work.

The quality indicators produced for the LFS are described below. Section 8.1 presents indicators related to sampling errors. Indicators related to non-sampling errors are discussed in Section 8.2. Section 8.3 describes the committees monitoring various aspects of the LFS to ensure the quality of the data released. Section 8.4 informs users of other resources available regarding LFS data quality.

## 8.1 Quality indicators related to sampling errors

Sampling error was defined earlier as the error that results from estimating a population parameter by measuring a portion of the population, the sample, rather than the entire population. The effect sampling errors have on survey estimates depends on several factors including the sample size, the sample design, the estimation method and the variability of the characteristic of interest.

If all other factors are constant, the sampling error is expected to decrease as the sample size increases. This is consistent with the fact that the sampling error should become zero once the entire population is sampled. For a given sample size, the sampling error is linked to the relative efficiency of various design characteristics. The stratification, the allocation and the selection method at each stage all have some impact on the magnitude

of the sampling error. The estimation method used also plays an important role for a given sample design. For example, the composite estimation method used by the LFS significantly reduces the sampling errors (See Chapter 6).

Finally, sampling errors differ from one variable to another since the degree of variability differs from one variable to another. These errors are generally greater for relatively rare characteristics and when the characteristic of interest is not distributed evenly in the population. Therefore, although they are based on the same sample, unemployment estimates generally have a higher sampling error than employment estimates.

For probability sample surveys, like the LFS, methods exist to calculate sampling errors. The most commonly used measure to quantify sampling error is sampling variance. The methods used for variance estimation in the case of the LFS have been presented in Chapter 7.

Three key measurements are derived from the sampling variance: the standard error (SE), the coefficient of variation (CV) and the design effect.

### 8.1.1 Standard error

The standard error, defined as the square root of the sampling variance, can be used to calculate a confidence interval associated with an estimate. The confidence interval is built around the resulting estimate and its width depends on the standard error and on a confidence level parameter.

To illustrate, the following example will be considered. In May 2015, the LFS estimate for the unemployment rate of the Canadian population 15 years of age and up was 6.8%, and the standard error associated with this estimate was 0.001395. An approximate 68% confidence interval for the true unemployment rate is then given by 0.068±1x(0.001395), or between 6.66% and 6.94%. The confidence level means that if the same selection and estimation process was repeated several times (leading to different samples and different estimates), 68% of the confidence intervals built this way would contain the true population value.

The estimates of change from one month to the next have become more important to users over time. In response, the monthly LFS release now provides the standard errors (SEs) for the provincial and national month-to-month changes for employed and unemployed.

Given their stability, the SEs included in the monthly LFS publication are not updated every month. Instead, an estimate of the SE that corresponds to the average of the SEs from the twelve previous months is provided. These estimates are updated twice a year (usually in January and July). The table below provides the SEs observed for the month-to-month change in employment and unemployment estimates for Canadians 15 years of age and up.

**Table 8.1**
**Standard error (SE) of the variation from one month to the next, Employed and Unemployed**

| | Employed | Unemployed |
|---|---|---|
| Province | thousands | |
| Newfoundland and Labrador | 2.1 | 2.1 |
| Prince Edward Island | 0.6 | 0.6 |
| Nova Scotia | 2.7 | 2.5 |
| New Brunswick | 2.3 | 2.1 |
| Quebec | 15.9 | 13.7 |
| Ontario | 19.3 | 17.0 |
| Manitoba | 2.6 | 2.1 |
| Saskatchewan | 2.7 | 2.1 |
| Alberta | 9.8 | 8.1 |
| British Columbia | 10.6 | 8.5 |
| Canada | 29.5 | 25.3 |

### 8.1.2 Coefficient of variation

The Coefficient of Variation (CV), which is defined as the standard error divided by the estimate, is a relative measure of variation and is usually expressed as a percentage. In the example used earlier, the CV for the May

2015 unemployment rate is 2.05% ((0.001395/0.068)x100%). It gives an indication of the uncertainty associated with the estimates. Small CVs are desirable because they indicate that the sampling variability is small relative to the estimate.

In order to obtain CVs, the users are provided with approximate CV tables. These tables give approximate CVs according to the observed values of the estimates, for various domains. The values are conservative in the sense that, if many survey estimates were to be produced for the same domain, around 75% of the approximated CVs obtained from the tables will be larger than the actual CVs that would be calculated if the precise methods were used. There will, however, be 25% of approximated CVs that will be somewhat lower than the precise calculation. This has the net effect of producing quality indicators that show lower quality of the survey estimates than is actually the case – confidence intervals are wider and statistical tests show fewer significant differences. These approximate CV tables are updated annually and provided in the Guide to the Labour Force Survey (71-543-G).

### 8.1.3 Design effect

A third measure derived from the sampling variance is the design effect, a relative measure calculated by dividing the sampling variance of an estimate obtained under the survey design by the sampling variance of a Simple Random Sample (SRS) of the same sample size. It can also be used to compare the effectiveness of one sample design to another. In the case of the LFS, it is particularly useful as an indicator of the deterioration of the sample design over time, or as a comparison establishing the gain/loss in efficiency obtained by redesigning the survey or associated with modifying some components of the design.

Different types of design effects can be computed, and each one depends on the data used to establish it. Below, the term unadjusted design effect will be used to refer to design effects based on non-calibrated weights, meaning without the adjustment that takes the population counts and estimated totals into consideration. The term adjusted design effect will be used to refer to design effects that are based on the final weights, after composite calibration. As a result, the unadjusted design effects are indicative of the effectiveness of the sample design, while the adjusted design effects provide a more general evaluation of the overall strategy adopted by combining all the characteristics of the survey plan (stratification, multi-stage sampling, post-stratification and estimation). The smaller the design effect is, the more effective the design with regard to sampling variance. It should be noted that the unadjusted design effects (sample design) are generally greater than the adjusted design effects (survey plan) based on the final weights, since they do not benefit from the gain in precision from calibration.

The table below presents some values representing the averaged adjusted and unadjusted design effects for the characteristics employment and unemployment at the national and provincial levels, based on survey data from January to August 2015.

**Table 8.2**
**Design effects, Employed and Unemployed, 2015**

| | Employed | | Unemployed | |
|---|---|---|---|---|
| **Province** | Adjusted | Unadjusted | Adjusted | Unadjusted |
| Newfoundland and Labrador | 0.40 | 1.78 | 1.08 | 1.00 |
| Prince Edward Island | 0.31 | 1.28 | 1.00 | 1.03 |
| Nova Scotia | 0.35 | 1.85 | 1.08 | 1.17 |
| New Brunswick | 0.36 | 2.20 | 1.17 | 1.17 |
| Quebec | 0.50 | 2.70 | 1.66 | 1.96 |
| Ontario | 0.42 | 2.92 | 1.39 | 1.64 |
| Manitoba | 0.32 | 3.03 | 1.01 | 1.19 |
| Saskatchewan | 0.34 | 4.87 | 1.10 | 1.11 |
| Alberta | 0.48 | 4.25 | 1.44 | 1.66 |
| British Columbia | 0.44 | 3.52 | 1.44 | 1.59 |
| Canada | 0.54 | 3.73 | 1.77 | 2.08 |

In the LFS, unadjusted design effects, together with other information, are used to identify regions where the sample design has lost a significant portion of its effectiveness over time. In some cases, a mini-redesign is performed in these regions to remedy this problem.

## 8.2 Quality indicators related to non-sampling errors

Non-sampling errors are errors that arise during the course of virtually all survey activities, apart from sampling. The impact on the estimates can be seen in the bias and/or variability of the estimates. If these errors are random errors, then their effects will approximately cancel out over a large enough domain, leading solely to increased variability. However, the effect can still be large for small domains or when the characteristics being studied are rare. If the errors are systematic, in the sense that they tend to go in the same direction, this will lead to a bias in the final results. And unlike random errors, the bias linked to systematic errors cannot be reduced by increasing the size of the sample.

The most common sources of non-sampling errors are coverage error, nonresponse, measurement or response errors and processing errors. Each one is discussed separately in the following sections.

### 8.2.1 Coverage errors

Coverage errors consist of omissions, erroneous inclusions, duplications and misclassifications of units on the survey frame. In the case of the LFS, those errors may happen when the list of dwellings associated with a PSU is established or uploaded, when listing maintenance is performed to identify growth, when the dwellings and/or the persons to include in the survey are contacted, or when data are collected and processed. In the LFS, three main indicators are used to measure and monitor the coverage errors: the slippage rate, the vacancy rate and the PSU yield evaluation.

The slippage rate is the relative difference between the population size estimates produced from the pre-calibration weights and the most recent population projection estimates used as calibration totals.

The population projection estimates used to determine the slippage rate can also contain errors, and these errors are one of the factors that contribute to slippage. In the LFS, undercoverage is typically observed, as indicated by a positive slippage rate. To reduce the resulting bias as much as possible, the weight of each respondent is modified by the composite calibration adjustment factor (see Chapter 6).

Undercoverage is the result of omitting dwellings or persons from the target population. An occupied dwelling may not be on the PSU list for various reasons: it was omitted when the list was being established, the building was under construction when it was last verified, there were errors in the cluster delineations, or it was wrongly classified as vacant. It is also possible that persons in the household were overlooked, either because the respondent did not make their existence known or they were classified as being a member of a usual place of residence other than the dwelling sampled. Students are often overlooked since they live elsewhere during their studies, even though their usual residence is in the sample. Therefore, errors can slip into the survey estimates if the characteristics of the individuals not included in the survey differ from those of the individuals included. For example, if the survey does not include a part of the population that is young and highly mobile with higher unemployment rates than the population of the same age in the survey, the slippage biases the unemployment estimates downward.

Slippage is also affected by population growth and nonresponse adjustments. The population grows between redesigns, and usually in specific places and unevenly. The selected sample can over- or underestimate this growth, or accurately account for it. For instance, the selected PSUs in an area may experience no growth, but other PSUs on the frame in the same area could be facing significant growth. In such a case, growth would be underestimated by the selected sample, and if the projected population estimates are in line with the growth that is actually occurring, the slippage rates would become larger for that area.

The adjustments to account for nonresponse (see Chapters 5 and 6) can also influence slippage. For instance, if non-respondent households have fewer members but are represented in the sample, via imputation or nonresponse adjustment factors, by large households, this can affect the slippage rate.

Lastly, as mentioned earlier, the population estimates also play a role in slippage. The more accurate they are, the more informative the slippage rates are.

Every month, the slippage rates are thoroughly analyzed. They are produced monthly at the national (excluding the territories) and provincial levels and for 12 age-sex groups (15-19, 20-24, 25-29, 30-39, 40-54, 55+). The table below provides the average slippage rates for the 2015 calendar year.

**Table 8.3**
**Average slippage rates - Canada by age group and province, 2015**

| Canada | % |
|---|---|
| All ages | 11.7 |
| 15 to 19 | 8.2 |
| 20 to 24 | 21.3 |
| 25 to 29 | 21.3 |
| 30 to 39 | 16.3 |
| 40 to 54 | 9.8 |
| 55+ | 7.0 |
| Newfoundland and Labrador | 11.6 |
| Prince Edward Island | 16.2 |
| Nova Scotia | 12.3 |
| New Brunswick | 11.7 |
| Quebec | 8.6 |
| Ontario | 12.0 |
| Manitoba | 9.5 |
| Saskatchewan | 13.7 |
| Alberta | 15.1 |
| British Columbia | 12.8 |

Dwellings correctly identified as being vacant or invalid do not introduce a bias into the LFS estimates. However, the estimation variance is higher because the sample contains fewer valid households. The LFS interviewers return to selected vacant dwellings every month to interview any persons targeted by the survey who may have moved in since the previous month. Non-existent dwellings are simply removed from the survey frame. Special attention must be given when determining vacant dwellings since they have a direct influence on two other indicators. If a dwelling is coded as vacant but its occupants are just temporarily absent, the nonresponse rate produced for the LFS will be underestimated. Furthermore, the slippage rate will be overestimated since this wrongly coded dwelling should have been considered when determining the rate. It is therefore important for interviewers to do a thorough job when determining whether a dwelling is vacant, and therefore out of the scope of the survey, or quite simply occupied by a temporarily absent household, and therefore within the scope of the survey. Vacancy rates are also produced and monitored on a monthly basis.

The table below presents the average vacancy rates and the minimum and maximum values for 2015 at the provincial and national levels.

**Table 8.4**
**Vacancy rate (unweighted), Canada and the Provinces, 2015**

| | Average | Maximum | Minimum |
|---|---|---|---|
| Province | | % | |
| Newfoundland and Labrador | 15.4 | 16.4 | 14.5 |
| Prince Edward Island | 21.1 | 23.4 | 19.8 |
| Nova Scotia | 18.1 | 18.9 | 17.4 |
| New Brunswick | 17.1 | 18.1 | 16.2 |
| Quebec | 12.8 | 13.3 | 12.2 |
| Ontario | 11.5 | 11.8 | 11.2 |
| Manitoba | 14.1 | 15.5 | 12.0 |
| Saskatchewan | 14.3 | 15.8 | 12.6 |
| Alberta | 14.6 | 15.2 | 13.7 |
| British Columbia | 12.3 | 13.2 | 11.9 |
| Canada | 13.7 | 14.1 | 13.0 |

For this quality indicator, there is some variability observed between provinces. This is linked to the proportion of seasonal dwellings owned varying from province to province. Seasonal dwellings are always considered as vacant, because they are not the usual place of residence for any of the occupants.

The yields of the PSUs are monitored monthly to detect any large differences between the number of dwellings surveyed in the field and the number of dwellings anticipated by the sample design. As a result, any significant discrepancy, such as 50% (positive or negative), between a DUF extract and the survey field results is reviewed. First, all clusters with an unexpected count are brought to the attention of the unit in Ottawa responsible for

controlling the sample, which verifies the cluster boundaries and the expected number of dwellings. If the discrepancy cannot be explained at the central office, the cluster is sent to the regional office in question for an in-depth analysis. All the causes that explain the discrepancies are filed for future reference.

This control plays an important role, because if the sample size requires changes, it is vital to know which regions are undersampled or oversampled. In addition, the discrepancies recorded can turn out to be problems for the survey and taint the quality of the LFS data.

All of these indicators (slippage rate, vacancy rate, and PSU yield) serve to detect potential problems with the sample coverage and to assist in taking any necessary action. Examples of possible actions are to put together training tools for interviewers to increase their knowledge of the household composition rules, to distribute a newsletter explaining slippage or the concept of multiple dwellings, or establish a program to relist a certain number of PSUs considered to be growing.

### 8.2.2 Nonresponse

Every month during the survey week, the interviewers are busy determining which selected dwellings contain persons eligible for the survey. Dwellings are identified as ineligible for the survey month for the following reasons:

- dwellings outside the scope of the survey, meaning a dwelling occupied by no persons who are part of the target population, *e.g.*, only members of the Canadian Armed Forces;
- vacant dwellings, meaning dwellings that are unoccupied, seasonal, under construction;
- invalid dwellings, meaning dwellings that are demolished, turned into business locations, moved (*e.g.*, mobile home), abandoned, or were initially entered by mistake.

When a dwelling is identified as eligible for the survey, it is not always possible to do an interview. This is called household nonresponse and can occur due to any number of reasons such as: no one at home, temporary absence, interview impossible (inclement weather, unusual circumstances in the household, *etc.*), technical problems, or refusal.

The magnitude of the bias due to nonresponse is usually not known, but it is directly linked to the differences in characteristics between the groups of responding units and the groups of non-responding units. Since the effect of this bias grows as the nonresponse rate increases, efforts are made to maintain the response rate as high as possible during collection.

The table below presents the average nonresponse rates as well as the minimum and maximum rates for 2015.

**Table 8.5**
**Nonresponse rates (unweighted), Canada and the Provinces, 2015**

| Province | Average | Maximum | Minimum |
|---|---|---|---|
| | | % | |
| Newfoundland and Labrador | 11.2 | 13.0 | 9.9 |
| Prince Edward Island | 10.9 | 12.2 | 9.0 |
| Nova Scotia | 10.3 | 11.3 | 9.7 |
| New Brunswick | 11.4 | 12.5 | 10.5 |
| Quebec | 10.2 | 11.9 | 8.2 |
| Ontario | 13.9 | 15.4 | 12.6 |
| Manitoba | 11.7 | 12.8 | 10.3 |
| Saskatchewan | 11.9 | 12.8 | 11.1 |
| Alberta | 12.8 | 14.0 | 11.5 |
| British Columbia | 11.7 | 12.6 | 10.6 |
| Canada | 12.0 | 13.1 | 11.2 |

Every month, the LFS produces nonresponse rates by cause (simple refusal, no contact, temporary absence, technical problem or other reason) and also by collection mode. These rates are carefully analyzed to identify the major causes of the nonresponse and to make any necessary corrections.

Refusal rates for the LFS are usually very low, with monthly Canadian rates varying between 1% and 2%. The refusal rates are usually similar across provinces, but can dip as low as 0.5% or climb as high as 3%.

To a certain extent, the collection system makes it possible to get more information on the reason for refusal, and thus allows tracking of the changes in respondents' attitudes toward the survey over time.

### 8.2.3 Measurement or Response errors

Measurement or response errors can be the result of the questionnaire design, how the questions are formulated, the respondent's comprehension, the way the interview is conducted, or the general survey conditions. They can occur when the information is provided, received, or entered into the computer. However, with the computerized collection method, it is possible to reduce some of these errors, since some verification rules are integrated into the collection instrument and conflicts must be resolved during the interview. Nevertheless, the respondent may incorrectly interpret the question, not know the answer, or have forgotten or altered the facts for personal reasons. In addition, interviewers can unintentionally re-interpret responses. As in the other error categories, response errors may lead to an increase in the variance and/or the presence of bias.

The proxy responses provided by one household member when information is collected about another household member can also lead to response errors. However, those errors are considered preferable to the nonresponse errors that would have to be dealt with if responses were only accepted by the respondent for him or herself. Currently, about 60% of the LFS information is provided by proxy and this rate remains fairly stable through time.

In repeated surveys, in which the sample consists of a certain number of panels or rotation groups, the expected value of estimates varies slightly from one rotation group to another. This is called rotation group bias. With regard to the LFS, this bias is at its highest level for the sixth of the sample in its first interview. It is possible to calculate the rotation effect by taking the ratio between an estimate calculated for the part of the sample participating in the survey a certain number of times (first month, second month, *etc*.) and the estimate calculated for the entire sample.

Brisebois and Mantel (1996) calculated a modified rotation effect that takes into account the differences in the effects of sampling errors for the six rotation groups. Their study revealed several statistically significant differences between the rotation groups, but the overall effect was determined to be minor.

### 8.2.4 Processing errors

Processing errors can occur at various stages of the survey, such as input, validation, verification, coding, imputation, weighting and data tabulation.

The computerized collection method helps to prevent skip errors during data input, since the application determines the flow of questions. Similarly, certain verification rules are integrated into the collection system to detect and correct discrepancies at the time of the interview.

The variables "occupation" and "industry" are coded to classification standards at the central office. In the first month of interviews, the interviewer collects information that accurately describes the type of company, industry or service in which the person works and that clearly and accurately indicates the type of work or nature of his/her duties. The first type of information is used to determine the industry, while the second type serves to identify the occupation. One of the first processing steps at the central office consists of coding the descriptive information collected for the variables "occupation" and "industry" based on the standard classification for these variables, NOC and NAICS. Monthly quality control processes are in place to evaluate the precision of this coding process.

The imputation rate is also a quality indicator with regard to data processing. Every month, diagnostics evaluating the results of the imputation process are produced and carefully examined. The diagnostics give information on the number of records treated by each imputation method and at each level of collapsing (see Chapter 5). The respective profiles of the non-imputed records and of the imputed ones are compared, as well as their respective contribution to the survey key estimates. This makes it possible to control the imputation quality and take the necessary actions.

To avoid errors likely to occur at the estimation and tabulation steps, a pre-release evaluation tool has been built. With the help of this tool, it is possible to highlight variables, subgroups and/or domains for which the estimates and/or the standard errors are unusually distant from their respective historical averages. These estimates can then be more specifically investigated to see if any error is responsible for the sudden change. In addition to

this, comparisons with other data sources are performed regularly, to verify if the LFS data is in line with other economic developments.

### 8.2.5 Monitoring of collection procedures

The collection application produces paradata files that contain a host of information on the activities of interviewers in the field and in call centres. Using these files, it is possible to produce quality indicators on the interviewers' activities. The LFS regularly analyzes the calls and visits made by the interviewers. Reports produced include, among others, information on the duration of the interviews (in person and over the telephone), the number of attempts to reach a respondent, and the number of cases transferred from one collection mode to another. Using this source of information, it is relatively easy to check whether the interviewers strictly follow the collection procedures and to take action in questionable cases These indicators can also be used to improve the training program for interviewers and strengthen certain components, such as task planning or the work schedule.

## 8.3 LFS committees

The LFS needs several coordination groups to see that the survey runs smoothly. Two permanent committees are described below. Their mandates include looking after permanent operations and evaluating the survey on a regular basis.

### 8.3.1 Operations Committee

The mandate of this committee is to monitor the activities that occur during each survey month and the circumstances surrounding the conduct of the survey, to ensure that the operations run smoothly, and to examine proposed changes and recommend whether they should be adopted. The Operations Committee is chaired by a senior member of Labour Statistics Division and meets every week.

### 8.3.2 Data Quality Committee

The committee, which was officially created in the spring of 1972, has the mandate of examining, evaluating and documenting monthly survey quality, and advising on any aspect of quality that needs attention. It also initiates and reviews ad hoc studies and investigations related to methods and procedures affecting data quality, and makes recommendations based on its findings. This committee is chaired by a member of the Household Survey Methods Division.

To ensure the best data quality possible, the Data Quality Committee periodically examines the different quality indicators described earlier. It meets every month to examine and assess the quality of the monthly data and to make suggestions and recommendations on any survey aspect likely to improve quality. By closely following the evolution of the quality indicators, the committee can intervene immediately with those in charge of the LFS activities in question to control the quality of the monthly data. The committee also discusses new developments that are likely to influence the quality of data that has just been collected or will be collected in the future, especially changes to the collection methods or the questionnaire, unusual problems in the field, ongoing testing of processes and methods, *etc*.

## 8.4 Resources available regarding LFS data quality

There are multiple other resources with information on various aspects of LFS data quality. This section will describe a few of them.

### 8.4.1 The Daily

The Labour Force Survey measures the current state of the Canadian labour market. Thanks to the data collected by the LFS, it is possible to produce various types of estimates (monthly estimate, estimate of change from one month to the next, three month moving average, *etc*.) for many different characteristics (labour force status, hours worked, multiple job holders, *etc*.), over thousands of domains (national, provincial, subprovincial, age-sex groups, *etc*.). Statistics Canada publishes the LFS estimates on a monthly basis, only ten days after the end of collection.

The publication of new LFS estimates, which usually takes place on the first Friday of the month, is announced through *The Daily*, Statistics Canada's official release bulletin, and is accompanied by a short analysis of the current labour market. The press release includes information on specific aspects of the survey, such as upcoming revisions, newly available reports and products, date of next release.

### 8.4.2 The Labour Force Survey web page

The Labour Force Survey web page, on Statistics Canada web site, has detailed information on many aspects of the survey, including quality. In particular, it contains information on the quality evaluation process and the various sources of data to which the LFS estimates are compared to see if labour market trends are in line with general economic performance. It also features a summary of the changes that occurred to the data or to the estimates through the years.

### 8.4.3 The Guide to the Labour Force Survey

The Guide to the Labour Force Survey (71-543-G) is a valuable source of information on survey concepts, classifications and definitions. It also provides guidelines and assistance on the comparison of LFS estimates across surveys (such as with the Survey of Employment, Payrolls and Hours) or across countries (such as with the USA). Appendix C also contains the Labour Force Survey questionnaire.

### 8.4.4 Access to LFS data

For users interested in the most common LFS estimates, the CANSIM tables are likely to have the information that is sought. Various types of estimates are provided for various domains and disclosure rules are applied to protect confidentiality.

For more specific situations, users may want to use the monthly released Public Use Micro-data File (71M0001X). This product is for users who prefer to do their own analysis and allows them to focus on specific subgroups in the population or cross-classify variables that are not in the catalogued products. Users can then submit requests on a cost-recovery basis to obtain variance estimates associated with their particular needs.

A Research Data Centre (RDC) provides access to Statistics Canada's confidential microdata files. They are accessible only to researchers with approved projects who have been sworn in as "deemed employees" of Statistics Canada. The RDC confidential microdata files contain most of the original information collected during the survey interview with the subject as well as derived variables added to the dataset afterwards. They also contain the bootstrap weights used to calculate the variance estimates, which are available only in the Master file. RDCs are located throughout the country. The following web site has more information:
www.statcan.gc.ca/eng/rdc/index.

The Real Time Remote Access (RTRA) complements existing methods of access to confidential micro-data. Using a secure username and password, the RTRA provides around the clock access to survey results from any computer with internet access. Confidentiality of the micro data is automated in the RTRA system, eliminating the need for manual intervention and allowing for rapid access to results. In order to utilize the RTRA Program, applicants must complete an application form. More information is provided on this web page: www.statcan.gc.ca/eng/rtra/rtra.

# Chapter 9 Using the LFS frame or sample for other surveys

## 9.0 Introduction

Many household surveys use the Labour Force Survey frame or sample for their survey design. Section 9.1 describes how the LFS frame is used by some other surveys to ensure coordination with the LFS. Section 9.2 describes how the LFS sample is used to obtain samples for supplementary or rotate-out surveys. Section 9.3 provides examples. Surveys that use the LFS frame or sample are important parts of the Statistics Canada household surveys program and are often sponsored by other government departments.

## 9.1 Surveys that use the LFS frame

Some surveys use the LFS frame to select a separate sample of households, usually in Primary Sampling Units (PSUs) that are also active in the LFS. Each survey reserves a set of random starts to select dwellings for their exclusive use. Based on the desired allocation, each stratum in the LFS may have zero, one or more starts reserved in this manner. In some cases, PSUs that will not be active for years in the LFS may also have random starts reserved for other surveys. Samples are selected using these starts. If the sample does not require full starts, the survey can do its own stabilization. This strategy of selecting separate samples reduces the respondent burden because it ensures that a dwelling cannot be selected by more than one survey. This is usually referred to as negative coordination of selected dwellings.

Although separate samples are selected for the other surveys that use the LFS frame, they can often share interviewer resources with the LFS since they are usually in the same PSUs. Sampling dwellings in the same area during the same collection period leads to collection cost reductions, especially for surveys with a high proportion of Computer Assisted Personal Interviews (CAPI). This strategy of selecting dwellings for different surveys from the same PSUs is referred to as positive coordination of selected PSUs.

## 9.2 Surveys that use the LFS sample

There are two types of surveys that use the LFS sample: supplementary and rotate-out surveys. Supplementary surveys interview households that have also been selected for the LFS and that are still active in the LFS. Dependent supplements use the LFS households while they are still being interviewed for the LFS, whereas independent supplements break off from the LFS to interview the LFS households at a separate time, or to allow more time than the LFS would for data collection. Rotate-out surveys are similar to supplementary surveys but contact the household after it has been rotated out of the LFS sample, *i.e.*, once the household has completed its sixth month of participation in the LFS.

The main advantage of supplementary and rotate-out surveys is that they can use the data collected by the LFS to screen respondents according to the survey needs. This can represent significant savings for surveys trying to reach households or persons with specific characteristics (*e.g.*, unemployed individuals). When a household has been rotated out of the LFS sample, that household is still eligible for rotate-out surveys for up to two years.

The primary concern with supplementary and rotate-out surveys is the respondent burden. Topics or questions that are likely to be unacceptable to respondents, or that could in some way influence responses obtained for the LFS in the following month, are avoided. Depending on the subject matter and/or the number of active surveys in a month, some supplements are well-received; they increase interviewing time, but on the other hand, they also add variety to the experience of being included in the LFS sample for six months.

Each of the six rotation groups of the LFS can be used to produce estimates. Typically, these surveys use one to five rotation groups for their sample, depending on the required level of reliability. For supplements, the LFS birth rotation group, *i.e.*, the one consisting of households being interviewed by the LFS for the first time, is usually avoided because of respondent burden. The initial LFS interview takes longer to complete than subsequent interviews.

In some cases, only some of a rotation group's households are required. Dwellings are dropped at random to reduce the sample to the required number of households, as in the LFS stabilization program. Within a selected dwelling, the survey may be directed at all eligible LFS respondents or at specific individuals. Separate individual

respondents may be selected from within selected dwellings through random selection or by screening for respondents with specific demographic or labour force characteristics from the LFS or through special questions.

Two other users of the active LFS sample are the Fast Track Option (FTO) module and the Disaster/Catastrophe Effects (DCE) module. The FTO module involves the addition of a small number of questions (no more than five) to the LFS questionnaire. Questions on a specific ad hoc topic are added for a single month and are asked to all respondents. This allows the survey to be conducted and the results to be released in a very timely manner. The DCE module is used to measure the economic impact of event such as natural disasters. In areas affected by such events, four questions are added to measure the number of working hours lost and the number of overtime hours worked. This module has been used, for example, to measure the economic impact of the 2013 floods in the Calgary region.

## 9.3 Examples of surveys that use the LFS frame or sample

The following list shows some of the surveys that used the LFS frame or sample in 2015.

**Table 9.1**
**Surveys that used the LFS frame or sample in 2015**

| Survey | Data collection period -2015 |
|---|---|
| **Supplementary surveys** | |
| Travel Survey of Residents of Canada (TSRC) | January to December (monthly) |
| Canadian Income Survey (CIS) | January to April |
| Elections Canada Fast Track Option | November |
| **Rotate-out surveys** | |
| Employment Insurance Coverage Survey (EICS) | April-May, July-August, November-December and January 2016 –February 2016 |
| **Surveys that use the LFS frame** | |
| Canadian Community Health Survey (CCHS) | January to December (4 quarterly collection periods) |
| Survey of Household Spending (SHS) | January to December (monthly) |

### 9.3.1 Canadian Community Health Survey

CCHS is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. It relies upon a large sample of respondents and is designed to provide reliable estimates at the health region level. The primary use of the CCHS data is for health surveillance and population health research. Federal and provincial departments of health and human resources, social service agencies, and other types of government agencies use the information collected from respondents to monitor, plan, implement and evaluate programs to improve the health of Canadians. The sample is divided on a yearly basis into four non-overlapping three-month collection periods. A large portion of the CCHS sample is positively coordinated with the sample of the LFS in terms of selected PSUs. This means that each month, the CCHS collection takes place in many PSUs in which the LFS is conducting CAPI interviews.

### 9.3.2 Survey of Household Spending

SHS is an annual survey that primarily collects detailed information on household expenditures. The SHS combines a questionnaire with recall periods based on the type of expenditure (1, 3 or 12 months, last payment, four weeks) and a daily expenditure diary that selected households complete for two weeks following the interview. SHS data are used at Statistics Canada by the System of National Accounts, in particular as input to calculate the gross domestic product (GDP). The data also helps to update the proportions (weights) of the Consumer Price Index (CPI). The data are collected on a continuous basis from January to December of the survey year, from a sample of households spread over twelve monthly collection cycles. Since 2015, most of the SHS sample is positively coordinated with the LFS sample in terms of selected PSUs. This means that each month, the SHS collection takes place almost exclusively in PSUs in which the LFS is conducting birth interviews.

### 9.3.3 Travel Survey of Residents of Canada

TSRC is sponsored by Statistics Canada, the Canadian Tourism Commission, and the provincial governments. It measures the size of domestic travel in Canada from the demand side. The objectives of the survey are to provide information about the volume of trips and expenditures for Canadian residents, to provide information on travel incidence and to provide the socio-demographic profile of travelers and non-travelers. It is a voluntary supplementary survey conducted monthly among LFS responding households that are in their second month of participation. Once the LFS interview is completed, a person 18 years of age or older is randomly selected from among the household members and the selected person is asked to answer the TSRC questionnaire.

### 9.3.4 Canadian Income Survey

The primary objective of CIS is to provide information on the income and income sources of Canadians, along with their individual and household characteristics. The survey gathers information on labour market activity, school attendance, disability, support payments, child care expenses, inter-household transfers, personal income, and characteristics and costs of housing. It is a supplementary survey conducted from January to April among LFS responding households that are in their sixth month of participation. Following the LFS interview, and subject to operational constraints, the interviewer asks the household member that provided the information for the LFS to answer the CIS questionnaire for all household members aged 16 years or older.

### 9.3.5 Employment Insurance Coverage Survey

The main purpose of EICS is to study the coverage of the employment insurance program. It provides a meaningful picture of who does or does not have access to EI benefits among the jobless and those in a situation of underemployment. The Employment Insurance Coverage Survey also covers access to maternity and parental benefits. The EICS is a rotate-out survey that uses the rotation groups that completed their sixth month in the LFS in March, June, October or December. Mothers from four additional rotation groups (one per collection cycle) are also selected to obtain an adequate sample size. There are four collection cycles each year: April-May, July-August, November-December and January-February. Each cycle lasts five weeks and begins during the month following the reference month (the last month of participation in the LFS).

## References

Address Register and Geography Methods Section (2015). Household Surveys Frame Service User Guide. *Social Survey Methods Division, Statistics Canada, Internal document.*

Alexander, C.H., Ernst, L.R. and Haas, M.E. (1982). A system for replacing primary sampling units when the units have been exhausted. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 211-216.

Benhin, E. and Mantel, H. (2012). Proposed Variance Estimation Method for the Labour Force Survey – from Jackknife to Bootstrap. *Paper presented at Statistics Canada's Advisory Committee on Statistical Methods* October 2012.

Bocci, C., and Beaumont, J.-F. (2004). Longitudinal Hot-deck Imputation for Household Nonresponse in the LFS. *Household Survey Methods Division, Statistics Canada, Internal document.*

Bocci, C., and Beaumont, J.-F. (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey. *Household Survey Methods Division, Statistics Canada, Internal document.*

Brisebois, F. and Mantel, H. (1996). Month-in-sample effects for the Canadian Labour Force Survey. *Proceedings of the Survey Methods Section,* Statistical Society of Canada.

Chen, E.J., and Liu, T.P. (2002). Choices of Alpha Value in Regression Composite Estimation for the Canadian Labour Force Survey: Impacts and Evaluation. *Methodology Branch Working Paper,* HSMD 2002-05E, Statistics Canada.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition, New York: John Wiley and Sons, Inc.

Drew, J.D., Choudhry, G.H. and Gray, G.B. (1978). Some methods for updating sample survey frames and their effects on estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Assocation, 62-71.

Fuller, W.A., and Rao, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.

Gagné, C., Roberts, G., and Keown, L.-A. (2014). Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software. *The Research Data Centres Information and Technical Bulletin*, Winter 2014, vol. 6 no. 1, Statistics Canada, Catalogue no. 12-002-X.

Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation. *Survey Methodology*, 27, 65-74.

Gray, G. (1973). Rotation of PSUs. *Statistics Canada, internal document.*

Kennedy, B. (1998). Weighting and Estimation Methodology of the Canadian Labour Force Survey. *Methodology Branch Working Paper* HSMD-98-002E, Statistics Canada.

Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *Journal of the American Statistical Association*, 46, 105-108.

Laflamme, G. (2003). Sélection et rotation des UPE : quelques précisions concernant la RAM, *Household Survey Methods Division, Statistics Canada, Internal document.*

Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology,* 13, 199-207.

Lorenz, P. (1996). Head Office Hot Deck Imputation System Specifications, Version 3. *Household Survey Methods Division, Statistics Canada, Internal document.*

Neusy, E. (2013). Adapting the Coordinated Bootstrap Method in the Presence of Change Between Survey Cycles. *Household Survey Methods Division, Statistics Canada, Internal document.*

Pandey, S., Alavi, A., and Beaumont, J.-F. (2003). Comparison of Integrated and Non-integrated Estimation Methods for GREG and Composite Estimators. *Household Survey Methods Division, Statistics Canada, Internal document.*

Phillips, O. (2004). Using Bootstrap Weights with WesVar and SUDAAN. *The Research Data Centres Information and Technical Bulletin*, Chronological index, Fall 2004, vol.1 no. 2, Statistics Canada, Catalogue no. 12-002-XIE.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). A simple procedure for unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 24, 482-491.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. *Journal of American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

*Road Network File, Reference Guide, 2015*. Statistics Canada Catalogue no. 92-500-G.

Roberts, G., Kovacevic, M., Mantel, H. and Phillips, O. (2001). Cross-sectional Inference Based on Longitudinal Surveys: Some Experiences with Statistics Canada Surveys. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, November 14-16, 2001.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New-York: Springer Verlag.

Singh, A.C., Kennedy, B., and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology,* 27, 33-44.

Statistics Canada (2003). *Survey Methods and Practices*. Statistics Canada, Catalogue no. 12-587-X.

Statistics Canada (2012). *Census Dictionary*. Statistics Canada, Catalogue no. 98-301-X.

White, S., and Benhin, E. (2013). Suggested Improvements for the LFS Person-Level Imputation. *Household Survey Methods Division, Statistics Canada, Internal document.*

# Appendix A.1 Glossary

**More detailed information on many of the terms in this glossary can be found at the following links.**

*Census Dictionary:*
http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/98-301-X2011001-eng.pdf

*Geography Catalogue:*
http://www.statcan.gc.ca/pub/92-196-x/92-196-x2011001-eng.pdf

*Guide to the Labour Force Survey:*
http://www.statcan.gc.ca/pub/71-543-g/71-543-g2015001-eng.pdf

**Address Register (AR)**
The Address Register is a database of residential addresses maintained to support the census and other household surveys. The database contains over 15,000,000 addresses and has national coverage, although it is more accurate in population centres. The three main products derived from the AR are the Dwelling Universe File (DUF), the Residential Telephone File (RTF) and the Socio-Economic File (SEF).

**Address Register (AR) Group**
The Dwelling Universe File extract used by LFS has PSU identification and sequence number applied to every dwelling. The completeness and accuracy of the extract at the PSU level is estimated in order to assign the PSU one of three group numbers.

- 0: PSUs in mailout area with no intial listing. The list of dwellings is considered to be of good quality. Selection is performed from the current list. Paper maps for listing maintenance are only generated upon request.
- 1: PSUs in the non-mailout area with no initial listing. The list of dwellings is considered to be of good quality. Selection is performed from the current list. Paper maps are generated automatically and may be used for list maintenance through time.
- 2: PSUs in the non-mailout area with initial listing. The quality of the list is considered of poor quality, and initial listing takes place before the first sample selection occurs. Paper maps are generated automatically and may be used for list maintenance through time.

**Allocation**
Allocation is the process of portioning out a fixed sample size into various provincial and/or subprovincial areas in order to satisfy various constraints on collection costs and/or on reliability of estimates.

**Area frame**
A frame of units based on geographical areas, such as Dissemination Areas or similar geography. It is usually used when there is not an adequate list frame of the ultimate survey units.

**Block**
A block, sometimes called a Census block or a Dissemination block, is an area bounded on all sides by roads and/or boundaries of standard geographic areas. Census blocks cover all the territory of Canada. They are the smallest geographic area for which population and dwelling counts are stored. It must be noted that blocks are not available to the public, but are used internally as the smallest level of geography upon which both collection and dissemination geographies are built.

**Block-face**
A block-face is one side of a street between two consecutive features intersecting that street. The features can be other streets or boundaries of standard geographic areas. Block-faces are used for generating block-face representative points, which in turn are used for geocoding and census data extraction when the street and address information are available.

**Collective dwelling**
A collective dwelling refers to a dwelling of a commercial, institutional or communal nature where people can reside but where the concept of a single family dwelling is difficult to apply. It includes lodging or rooming houses, hotels, motels, tourist homes, nursing homes, hospitals, staff residences, communal quarters (military bases), work camps, jails, group homes, and so on. Collective dwellings may be occupied by usual residents or solely by foreign residents and/or by temporarily present persons.

**Census division**
Census division (CD) is the general term for provincially legislated areas (such as county, *municipalité régionale de comté* and regional district) or their equivalents. Census divisions are intermediate geographic areas between the province/territory level and the municipality (census subdivision).

**Census metropolitan area or census agglomeration**
A census metropolitan area (CMA) or a census agglomeration (CA) is formed by one or more adjacent municipalities centred on a large urban area (known as the urban core). A CMA must have a total population of at least 100,000 of which 50,000 or more must live in the urban core. A CA must have an urban core population of at least 10,000. To be included in the CMA or CA, other adjacent municipalities must have a high degree of integration with the central urban area, as measured by commuting flows derived from census place of work data.

**Dissemination area**
A dissemination area (DA) is a small, relatively stable geographic unit composed of one or more adjacent dissemination blocks. It is the smallest standard geographic area for which all census data are disseminated. DAs cover all the territory of Canada. It is created as a unit for dissemination of Census data.

**Design effect**
The design effect of an estimator is the ratio of the actual variance of an estimator under the current sample design to what it would be under a simple random sampling design of the same number of elements.

**Dwelling**
Refers to a set of living quarters in which a person or a group of persons resides or could reside. Unoccupied dwellings are called vacant. For the LFS, a dwelling consists of any set of living quarters that is structurally separate and has a private entrance outside the building or from a common hall or stairway inside the building.

**Economic region**
An Economic region (ER) is a grouping of complete Census divisions (CDs) (with one exception in Ontario) created as a standard geographic unit for analysis of regional economic activity. They have been established in consultation with the provinces, except for Quebec, where Economic regions are designated by law (*'les régions administratives'*). ERs generally correspond to regions used by the province for administrative and statistical purposes. The boundaries in current use are based on the 2011 Standard Geographical Classification.

**Employment insurance economic region**
A set of regions across the country defined by Employment and Social Development Canada (ESDC) for the purpose of distributing Employment Insurance benefits in an equitable manner. The LFS is responsible for producing timely estimates required by ESDC in order to establish standards for admissibility to the program and the duration of benefits.

**Employment**
Employed persons are those who,during the reference week:

a. did any work at all at a job or business, that is, paid work in the context of an employer-employee relationship, or self-employment. It also includes persons who did unpaid family work, which is defined as unpaid work contributing directly to the operation of a farm, business or professional practice owned and operated by a related member of the same household; or
b. had a job but were not at work due to factors such as their own illness or disability, personal or family responsibilities, vacation or a labour dispute. This category excludes persons not at work because they were on layoff or between casual jobs, and those who did not then have a job (even if they had a job to start at a future date).

**Employment rate (employment population ratio)**
Number of employed persons expressed as a percentage of the population 15 years of age and over. The employment rate for a particular group (age, sex, marital status, province, etc.) is the number employed in that group expressed as a percentage of the population for that group.

**Household**
Any person or group of persons living in a dwelling. A household may consist of any combination of: one person living alone, one or more families, a group of people who are not related but who share the same dwelling. Note that foreign residents and persons with a usual place of residence elsewhere are not surveyed.

**Labour force**
Civilian non-institutional population 15 years of age and over who, during the survey reference week, were either employed or unemployed. Prior to 1966, only persons aged 14 and over were covered by the survey.

**Labour force status**
A labour force status classification (including employed, unemployed, and not in the labour force) is assigned to each respondent aged 15 and over, according to their responses to a number of questions during the interview.

**Listing**
Listing is the process by which the dwellings that belong to one area (usually a PSU) are recorded on paper or electronically. This field exercise is required for selected PSUs when a high quality list of dwellings is not available to perform the second stage selection of dwellings. Maps of the area, with clear boundaries, are required to determine where to list. Most listing is sequenced in a specific pattern in order to ensure all block-faces are examined, and in order to be able to re-locate a particular address months or years after initial listing.

**Listing maintenance**
Listing in the LFS proceeds in two stages. The second stage is ongoing maintenance of a pre-existing list. The list was originally generated by initial listing (AR group 2) or by initial loading (AR group 0 and AR group 1). The extent of changes is usually minor unless significant growth occurs in the area of the PSU, or significant errors are found in the initial listing effort or with the initial loading. Updates are sent directly to Head Office without involving senior interviewers, unless subsampling is requested due to significant growth.

**Multi-stage sampling**
Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called the primary sampling units (PSUs). The units selected at the second stage are called secondary sampling units (SSUs) or sometimes ultimate sampling units if it is the last stage. The units at each stage are different in structure and are hierarchical. In the case of the LFS, PSUs corresponds to dissemination areas selected within strata. In the second stage, dwellings are selected within each first-stage selected PSU.

**Participation rate**
The participation rate represents the labour force expressed as a percentage of the population 15 years of age and over. The participation rate for a particular group (age, sex, etc.) is the labour force in that group expressed as a percentage of the population for that group.

**Population centre**
A population centre has a population of at least 1,000 and a population density of 400 persons or more per square kilometre, based on the current census population count. All areas outside population centres are classified as rural areas. Taken together, population centres and rural areas cover all of Canada. Population centre population includes all population living in the cores, secondary cores and fringes of Census metropolitan areas (CMAs) and Census agglomerations (CAs), as well as the population living in population centres outside CMAs and CAs.

**Primary sampling unit**
Units selected at the first stage of sampling in a multistage design are called primary sampling units, or PSUs. With the 2015 redesign, the LFS PSUs are mainly defined as dissemination areas.

**Probability proportional to size sampling**
Probability proportional to size (PPS) sampling is a technique that uses auxiliary information and yields unequal probability of inclusion. If population units vary in size and these sizes are known, such information can be used during sampling to increase the statistical efficiency. In the case of the LFS, the PSUs are selected with PPS, and the size measure used is related to the approximate number of households in each PSU. More information on the size measure is provided in Section 6.2.1.

**Reference period**
A period of time used in surveys for which respondents must recall and answer. For example, "how many hours did you work last week?" In the LFS, the reference week is usually the week containing the 15th day of the month.

**Rotation**
Sample rotation is the periodic replacement of one unit with another. The LFS has

- Dwelling rotation (within a PSU) after six months in the survey
- PSU rotation after two to fifty years in the survey, with an average around ten years. In many cases, there is a survey redesign before rotation of the PSU takes place.

The set of dwellings (or the PSUs that contain them) that rotate in the same month are referred to as a rotation panel or a rotation group. Each panel consists of one sixth of the sample. As a result, each month has a mix of dwellings in their first, second, third, fourth, fifth and sixth interview.

**Rural area**
Rural areas include all territory lying outside population centres. Taken together, population centres and rural areas cover all of Canada. Rural population includes all population living in the rural areas of Census metropolitan areas (CMAs) and Census agglomerations (CAs), as well as population living in rural areas outside CMAs and CAs.

**Sampling rate**
The sampling rate is the ratio of the size of the sample to the size of the population on the frame. A 1 in 20 sample would select 5% of the units for data collection and have a 0.05 sampling rate, or an inverse sampling rate of 20.

**Sampling variance**
The sampling variance measures the extent to which the estimates of a population parameter, obtained from all different possible samples selected using the same design, differ from one another. It is calculated as the average value of the squared difference of the estimate from its mean over all possible samples.

**Slippage rate**
The slippage rate is a measure of discrepancy between an estimate of the size of a population (e.g. province or age-sex group) and the corresponding Census-projected value for the same population. It equals (1 – (ratio of estimate to projection))x100%.

**Stratification**
Stratification is the process of dividing the survey primary sampling units into homogeneous, mutually exclusive groups called strata, and then samples are selected independently from each stratum.

**Systematic sampling**
Systematic sampling is a method of selecting a sample in which the first item is selected from the population randomly (random start), with the remaining sample items drawn at equally spaced intervals (according to the inverse sampling rate). With a sampling rate of one in ten, and a random starting point of seven, the 7$^{th}$ unit is selected, and every 10th unit thereafter is selected (17$^{th}$, 27$^{th}$, 37$^{th}$, *etc.*) until the end of the list is reached.

**Target population**
The target population is the population for which information is desired. The target population covered by the LFS corresponds to all persons aged 15 years and over residing in the provinces of Canada, with the exception of the following: persons living on Indian reserves, full-time members of the regular Armed Forces, and persons living in institutions (for example, inmates of penal institutions and patients in hospitals or nursing homes who have resided in the institution for more than six months).

**Three month moving average estimate**
A three month moving average estimate is an average of the values of the estimate for each of the most recent three months. It can be produced every month, using the most recent three months. The average is defined slightly differently whether the estimate is a total or a rate. Mathematical expression is given in Section 6.1.

**Unemployment**
Unemployed persons are those who, during the reference week:

a. were without work but had looked for work in the past four weeks ending with the reference period and were available for work; or

b. were on temporary layoff due to business conditions, with an expectation of recall, and were available for work; or

c. were without work, had a job to start within four weeks from the reference period and were available for work.

**Unemployment rate**
Number of unemployed persons expressed as a percentage of the labour force. The unemployment rate for a particular group (for example, age, sex, marital status) is the number of unemployed in that group expressed as a percentage of the labour force for that group. This rate is one of the key statistics produced by the LFS.

**Vacancy rate**
The vacancy rate is the proportion of unoccupied dwellings. Out-of-scope dwellings such as businesses and demolished dwellings are not included in the denominator.

## Appendix A.2 Abbreviations

| | |
|---|---|
| AR | Address Register |
| CA | Census agglomeration |
| CANSIM | Statistics Canada's key socioeconomic database |
| CAPI | Computer assisted personal interviewing |
| CATI | Computer assisted telephone interviewing |
| CAWI | Computer assisted web interviewing |
| CCHS | Canadian Community Health Survey |
| CIS | Canadian Income Survey |
| CMA | Census metropolitan area |
| CPI | Consumer Price Index |
| CSD | Census subdivision |
| CV | Coefficient of variation |
| DA | Dissemination area (census) |
| DCE | Disaster/Catastrophe Effects module |
| DUF | Dwelling Universe File |
| EICS | Employment Insurance Coverage Survey |
| EIER | Employment insurance economic region |
| EQ | Electronic questionnaire |
| ER | Economic Region |
| ESDC | Employment and Social Development Canada |
| FTO | Fast Track Option module |
| GDP | Gross Domestic Product |
| GMS | Generalized Mapping System |
| HDIS | Hot-Deck Imputation System |
| HOPS | Head Office Processing System |
| ISD | Interviewer Selected Dwellings |

| | |
|---|---|
| ISR | Inverse sampling ratio |
| LFS | Labour Force Survey |
| NAICS | North American Industry Classification System |
| NGD | National Geographic Database |
| NOC | National Occupational Classification |
| NOC-S | National Occupational Classification - Statistics |
| PPS | Probability proportional to size |
| PPSWR | Probability proportional to size with replacement |
| PPSWOR | Probability proportional to size without replacement |
| PSU | Primary sampling unit |
| RHC | Rao-Hartley-Cochran (random group method) |
| RTF | Residential Telephone File |
| SE | Standard Error |
| SHS | Survey of Household Spending |
| SRS | Simple Random Sampling |
| SSU | Secondary sampling unit |
| TFC | Telephone First Contact |
| TSRC | Travel Survey of Residents of Canada |

# Appendix B Characteristics of the survey frame and the sample design

**Table B.1**
**Number of households covered by the frame and provincial sample sizes**

| Province | Households covered by the frame | Households excluded from the frame | Sample financed by Statistics Canada | Sample financed by ESDC | Total sample |
|---|---|---|---|---|---|
| | number | | | | |
| Newfoundland and Labrador | 211,361 | 7,878 | 1,852 | 157 | 2,009 |
| Prince Edward Island | 59,402 | 277 | 1,421 | 0 | 1,421 |
| Nova Scotia | 408,853 | 4,281 | 2,965 | 0 | 2,965 |
| New Brunswick | 335,516 | 5,536 | 2,623 | 187 | 2,810 |
| Quebec | 3,620,953 | 23,468 | 5,108 | 5,077 | 10,185 |
| Ontario | 5,108,516 | 49,030 | 7,036 | 7,936 | 14,972 |
| Manitoba | 467,146 | 18,399 | 3,911 | 295 | 4,206 |
| Saskatchewan | 400,608 | 25,012 | 3,904 | 218 | 4,122 |
| Alberta | 1,487,737 | 25,183 | 3,690 | 810 | 4,500 |
| British Columbia | 1,881,831 | 57,057 | 3,507 | 1,920 | 5,427 |
| Canada | 13,981,923 | 216,121 | 36,017 | 16,600 | 52,617 |

**Table B.2**
**Number of households in remote strata, large-cluster strata, and one-stage strata, by province**

| | Remote strata | | Large-cluster strata | | One-stage strata | |
|---|---|---|---|---|---|---|
| Province | Strata | Households | Strata | Households | Strata | Households |
| | number | | | | | |
| Newfoundland and Labrador | 0 | 0 | 0 | 0 | 0 | 0 |
| Prince Edward Island | 0 | 0 | 0 | 0 | 138 | 59,402 |
| Nova Scotia | 0 | 0 | 0 | 0 | 0 | 0 |
| New Brunswick | 0 | 0 | 0 | 0 | 0 | 0 |
| Quebec | 0 | 0 | 0 | 0 | 0 | 0 |
| Ontario | 5 | 45,370 | 4 | 105,612 | 0 | 0 |
| Manitoba | 5 | 12,751 | 0 | 0 | 0 | 0 |
| Saskatchewan | 1 | 4,100 | 0 | 0 | 0 | 0 |
| Alberta | 1 | 3,847 | 0 | 0 | 0 | 0 |
| British Columbia | 5 | 45,366 | 0 | 0 | 0 | 0 |
| Canada | 17 | 111,434 | 4 | 105,612 | 138 | 59,402 |

**Table B.3.1**
**Statistics for the high-income strata**

| Census Metropolitan Area (CMA) | High-income strata | Households in high-income strata | Prevalence[1] of high-income households[2] in high-income strata | Prevalence of high-income households in the CMA | High-income households in the CMA that are in a high-income stratum |
|---|---|---|---|---|---|
| | number | | | % | |
| St. John's | 1 | 5,998 | 41.1 | 14.5 | 19.8 |
| Halifax | 2 | 14,173 | 36.1 | 11.0 | 26.4 |
| Saguenay | 1 | 5,436 | 22.8 | 6.7 | 24.7 |
| Quebec City | 1 | 20,720 | 34.9 | 8.1 | 24.2 |
| Sherbrooke | 1 | 6,224 | 23.7 | 5.7 | 26.2 |
| Trois-Rivières | 1 | 5,118 | 24.4 | 5.5 | 29.7 |
| Montreal | 3 | 117,935 | 37.2 | 9.1 | 28.1 |
| Ottawa-Gatineau | 2 | 44,015 | 48.4 | 17.5 | 22.8 |
| Kingston | 1 | 3,722 | 40.9 | 12.0 | 18.6 |
| Oshawa | 1 | 11,087 | 41.6 | 15.8 | 21.3 |
| Toronto | 6 | 159,618 | 51.2 | 16.9 | 23.0 |
| Hamilton | 1 | 17,896 | 47.1 | 13.8 | 20.6 |
| St. Catharines - Niagara | 1 | 12,805 | 28.3 | 8.9 | 24.1 |
| Kitchener - Waterloo | 1 | 12,927 | 45.0 | 13.0 | 23.4 |
| London | 1 | 13,367 | 40.8 | 10.6 | 25.3 |
| Windsor | 1 | 11,131 | 33.0 | 10.0 | 28.0 |
| Greater Sudbury | 1 | 4,201 | 41.7 | 13.1 | 18.3 |
| Thunder Bay | 1 | 3,300 | 35.6 | 10.8 | 20.2 |
| Winnipeg | 4 | 24,268 | 37.9 | 10.0 | 30.1 |
| Regina | 1 | 5,138 | 54.2 | 16.2 | 18.8 |
| Saskatoon | 2 | 9,023 | 45.1 | 14.6 | 25.3 |
| Calgary | 1 | 23,880 | 64.6 | 23.4 | 13.0 |
| Edmonton | 1 | 24,449 | 56.0 | 19.6 | 14.2 |
| Abbotsford | 1 | 4,845 | 28.2 | 10.4 | 20.4 |
| Vancouver | 3 | 66,687 | 40.5 | 14.3 | 19.4 |
| Victoria | 1 | 9,662 | 32.7 | 11.2 | 17.1 |

1. The reported prevalence is the prevalence according to the 2011 T1 Family File.
2. A high-income household is a household with a reported annual household income over $150,000.

**Table B.3.2**
**Statistics for the immigrant strata**

| Province | Immigrant strata | Households in the immigrant strata | Prevalence of Immigrant households[1] in the immigrant strata | Prevalence of Immigrant households in the province | Immigrant households in the province that are in an immigrant stratum |
|---|---|---|---|---|---|
| | number | | | % | |
| Manitoba | 2 | 11,861 | 16.1 | 6.3 | 6.5 |

1. An immigrant household is a household for which at least one member reported having immigrated to Canada in the last ten years, according to the 2011 National Household Survey.

**Table B.3.3**
**Statistics for the Aboriginal strata**

| Province | Aboriginal strata | Households in the Aboriginal strata | Prevalence of Aboriginal households[1] in the Aboriginal strata | Prevalence of Aboriginal households in the province | Aboriginal households in the province that are in an Aboriginal stratum |
|---|---|---|---|---|---|
| | number | | | % | |
| Saskatchewan | 3 | 13,743 | 36.5 | 10.9 | 11.5 |
| Alberta | 8 | 123,679 | 20.2 | 5.9 | 28.3 |
| British Columbia | 9 | 143,433 | 17.1 | 5.3 | 24.7 |

1. An Aboriginal household is a household in which at least one member reported having an Aboriginal status according to the 2011 National Household Survey.

**Table B.4**
**Characteristics of sample design by sub-provincial regions**

| Province | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| | number | | | | number | % | number | % |
| **Newfoundland and Labrador** | **995** | **37** | **493,776** | **241,034** | **211,361** | **100.0** | **2,009** | **100.0** |
| EIER | | | | | | | | |
| 01 | 371 | 14 | 194,504 | 90,920 | 84,973 | 40.2 | 692 | 34.4 |
| 02 | 624 | 23 | 299,272 | 150,114 | 126,388 | 59.8 | 1,317 | 65.6 |
| ER | | | | | | | | |
| 1010 | 512 | 19 | 259,803 | 124,902 | 112,432 | 53.2 | 971 | 48.3 |
| 1020 | 76 | 3 | 34,938 | 17,406 | 14,668 | 6.9 | 150 | 7.5 |
| 1030 | 192 | 7 | 95,199 | 45,846 | 39,773 | 18.8 | 433 | 21.6 |
| 1040 | 215 | 8 | 103,836 | 52,880 | 44,488 | 21.0 | 455 | 22.6 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| St. John's (CMA) | 375 | 14 | 196,966 | 91,949 | 85,892 | 40.6 | 699 | 34.8 |
| Corner Brook | 52 | 2 | 27,433 | 12,105 | 11,503 | 5.4 | 125 | 6.2 |
| Other Urban | 235 | 10 | 118,513 | 54,344 | 50,459 | 23.9 | 525 | 26.1 |
| Non-Urban | 333 | 11 | 150,864 | 82,636 | 63,507 | 30.0 | 659 | 32.8 |
| Stratum type | | | | | | | | |
| Regular | 968 | 36 | 477,623 | 234,782 | 205,363 | 97.2 | 1,960 | 97.6 |
| High income | 27 | 1 | 16,153 | 6,252 | 5,998 | 2.8 | 49 | 2.4 |
| **Prince Edward Island** | **NA** | **138** | **139,424** | **69,987** | **59,402** | **100.0** | **1,421** | **100.0** |
| EIER | | | | | | | | |
| 03 | NA | 66 | 64,438 | 31,227 | 28,392 | 47.8 | 679 | 47.8 |
| 62 | NA | 72 | 74,986 | 38,760 | 31,010 | 52.2 | 742 | 52.2 |
| ER | | | | | | | | |
| 1110 | NA | 138 | 139,424 | 69,987 | 59,402 | 100.0 | 1,421 | 100.0 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Charlottetown | NA | 66 | 64,438 | 31,227 | 28,392 | 47.8 | 679 | 47.8 |
| Summerside | NA | 18 | 16,488 | 7,747 | 7,342 | 12.4 | 176 | 12.4 |
| Other Urban | NA | 7 | 7,447 | 3,925 | 3,286 | 5.5 | 79 | 5.5 |
| Non-Urban | NA | 47 | 51,051 | 27,088 | 20,382 | 34.3 | 488 | 34.3 |
| Stratum type | | | | | | | | |
| One-stage | NA | 138 | 139,424 | 69,987 | 59,402 | 100.0 | 1,421 | 100.0 |
| **Nova Scotia** | **1,726** | **57** | **909,763** | **459,722** | **408,853** | **100.0** | **2,965** | **100.0** |
| EIER | | | | | | | | |
| 04 | 333 | 12 | 166,885 | 88,754 | 74,769 | 18.3 | 654 | 22.0 |
| 05 | 704 | 23 | 362,358 | 188,303 | 162,738 | 39.8 | 1,242 | 41.9 |
| 06 | 689 | 22 | 380,520 | 182,665 | 171,346 | 41.9 | 1,069 | 36.1 |
| ER | | | | | | | | |
| 1210 | 263 | 10 | 129,954 | 67,389 | 58,419 | 14.3 | 532 | 17.9 |
| 1220 | 304 | 11 | 152,348 | 81,041 | 69,226 | 16.9 | 546 | 18.4 |
| 1230 | 230 | 7 | 121,924 | 59,335 | 53,427 | 13.1 | 382 | 12.9 |
| 1240 | 224 | 7 | 115,445 | 63,490 | 52,170 | 12.8 | 409 | 13.8 |
| 1250 | 705 | 22 | 390,092 | 188,467 | 175,611 | 43.0 | 1,096 | 37.0 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Halifax (CMA) | 705 | 22 | 390,092 | 188,467 | 175,611 | 43.0 | 1,096 | 37.0 |
| Cape Breton | 195 | 8 | 97,398 | 48,831 | 44,326 | 10.8 | 404 | 13.6 |
| Truro | 84 | 3 | 45,041 | 23,038 | 20,688 | 5.1 | 163 | 5.5 |
| New Glasgow | 72 | 3 | 35,342 | 17,843 | 16,199 | 4.0 | 128 | 4.3 |
| Other Urban | 213 | 8 | 108,377 | 55,000 | 50,043 | 12.2 | 379 | 12.8 |
| Non-Urban | 457 | 13 | 233,513 | 126,543 | 101,986 | 24.9 | 795 | 26.8 |
| Stratum type | | | | | | | | |
| Regular | 1,657 | 55 | 868,980 | 445,069 | 394,680 | 96.5 | 2,877 | 97.0 |
| High income | 69 | 2 | 40,783 | 14,653 | 14,173 | 3.5 | 88 | 3.0 |

**Table B.4 (continued)**
**Characteristics of sample design by sub-provincial regions**

| Province | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| | number | | | | number | % | number | % |
| **New Brunswick** | **1,478** | **50** | **737,765** | **371,662** | **335,516** | **100.0** | **2,810** | **100.0** |
| EIER | | | | | | | | |
| 07 | 750 | 24 | 388,553 | 188,036 | 174,494 | 52.0 | 1,191 | 42.4 |
| 08 | 225 | 9 | 110,386 | 57,003 | 51,141 | 15.2 | 595 | 21.2 |
| 09 | 503 | 18 | 238,826 | 126,623 | 109,881 | 32.7 | 1,024 | 36.5 |
| ER | | | | | | | | |
| 1310 | 337 | 12 | 155,303 | 78,953 | 72,071 | 21.5 | 697 | 24.8 |
| 1320 | 398 | 12 | 201,789 | 103,053 | 91,456 | 27.3 | 657 | 23.4 |
| 1330 | 337 | 11 | 168,517 | 83,684 | 76,248 | 22.7 | 627 | 22.3 |
| 1340 | 247 | 9 | 133,660 | 66,365 | 59,181 | 17.6 | 440 | 15.7 |
| 1350 | 159 | 6 | 78,496 | 39,607 | 36,560 | 10.9 | 389 | 13.8 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Moncton (CMA) | 273 | 8 | 138,596 | 67,328 | 62,885 | 18.7 | 408 | 14.5 |
| Saint John (CMA) | 256 | 8 | 127,813 | 62,684 | 57,608 | 17.2 | 392 | 14.0 |
| Fredericton | 165 | 6 | 92,914 | 44,250 | 41,500 | 12.4 | 291 | 10.3 |
| Bathurst | 75 | 3 | 33,343 | 17,717 | 16,130 | 4.8 | 156 | 5.6 |
| Miramichi | 55 | 2 | 26,889 | 13,272 | 12,157 | 3.6 | 118 | 4.2 |
| Edmundston | 45 | 2 | 21,698 | 12,065 | 11,277 | 3.4 | 120 | 4.3 |
| Other Urban | 169 | 6 | 82,037 | 41,339 | 38,235 | 11.4 | 379 | 13.5 |
| Non-Urban | 440 | 15 | 214,475 | 113,007 | 95,724 | 28.5 | 947 | 33.7 |
| Stratum type | | | | | | | | |
| Regular | 1,478 | 50 | 737,765 | 371,662 | 335,516 | 100.0 | 2,810 | 100.0 |
| **Quebec** | **15,296** | **210** | **7,831,321** | **3,914,006** | **3,620,953** | **100.0** | **10,185** | **100.0** |
| EIER | | | | | | | | |
| 10 | 281 | 8 | 135,446 | 71,862 | 64,498 | 1.8 | 471 | 4.6 |
| 11 | 1,522 | 19 | 756,546 | 380,465 | 364,288 | 10.1 | 878 | 8.6 |
| 12 | 316 | 12 | 148,215 | 80,176 | 74,992 | 2.1 | 723 | 7.1 |
| 13 | 301 | 13 | 154,167 | 78,491 | 70,221 | 1.9 | 813 | 8.0 |
| 14 | 337 | 11 | 169,087 | 89,244 | 82,807 | 2.3 | 664 | 6.5 |
| 15 | 1,021 | 15 | 520,259 | 256,370 | 240,625 | 6.6 | 721 | 7.1 |
| 16 | 7,236 | 58 | 3,791,701 | 1,788,374 | 1,699,706 | 46.9 | 2,042 | 20.0 |
| 17 | 2,035 | 18 | 1,033,672 | 580,860 | 497,902 | 13.8 | 919 | 9.0 |
| 18 | 458 | 13 | 223,026 | 128,442 | 106,911 | 3.0 | 705 | 6.9 |
| 19 | 907 | 14 | 440,313 | 234,837 | 207,541 | 5.7 | 744 | 7.3 |
| 20 | 570 | 18 | 305,660 | 148,838 | 138,635 | 3.8 | 860 | 8.4 |
| 21 | 312 | 11 | 153,229 | 76,047 | 72,827 | 2.0 | 644 | 6.3 |
| ER | | | | | | | | |
| 2410 | 186 | 6 | 91,501 | 47,889 | 43,092 | 1.2 | 352 | 3.5 |
| 2415 | 424 | 7 | 199,492 | 108,381 | 96,591 | 2.7 | 332 | 3.3 |
| 2420 | 1,403 | 16 | 699,011 | 364,264 | 341,854 | 9.4 | 756 | 7.4 |
| 2425 | 810 | 18 | 410,444 | 204,018 | 188,150 | 5.2 | 1,053 | 10.3 |
| 2430 | 620 | 18 | 310,558 | 170,550 | 150,125 | 4.1 | 1,016 | 10 |
| 2433 | 484 | 5 | 233,940 | 117,487 | 110,318 | 3.0 | 247 | 2.4 |
| 2435 | 2,743 | 27 | 1,441,851 | 665,695 | 636,527 | 17.6 | 1,160 | 11.4 |
| 2440 | 3,682 | 32 | 1,886,479 | 964,739 | 902,456 | 24.9 | 1,162 | 11.4 |
| 2445 | 722 | 6 | 401,555 | 169,067 | 163,577 | 4.5 | 200 | 2.0 |
| 2450 | 903 | 6 | 468,048 | 222,711 | 203,375 | 5.6 | 239 | 2.3 |
| 2455 | 1,055 | 9 | 556,460 | 286,272 | 248,081 | 6.9 | 370 | 3.6 |
| 2460 | 693 | 20 | 366,212 | 189,489 | 167,141 | 4.6 | 1,004 | 9.9 |
| 2465 | 288 | 8 | 141,432 | 74,126 | 68,367 | 1.9 | 383 | 3.8 |
| 2470 | 532 | 13 | 259,237 | 144,550 | 131,193 | 3.6 | 764 | 7.5 |
| 2475 | 551 | 14 | 271,650 | 139,028 | 127,857 | 3.5 | 818 | 8.0 |
| 2480 | 167 | 4 | 79,621 | 39,055 | 36,210 | 1.0 | 232 | 2.3 |
| 2490 | 33 | 1 | 13,830 | 6,685 | 6,039 | 0.2 | 97 | 1.0 |

**Table B.4 (continued)**
**Characteristics of sample design by sub-provincial regions**

| | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| **Province** | number | | | | number | % | number | % |
| **CMA/CA-Design Type Indicator** | | | | | | | | |
| Saguenay (CMA) | 318 | 11 | 157,063 | 78,373 | 74,477 | 2.1 | 659 | 6.5 |
| Quebec (CMA) | 1,540 | 19 | 765,457 | 384,703 | 368,299 | 10.2 | 889 | 8.7 |
| Sherbrooke (CMA) | 409 | 14 | 201,890 | 108,641 | 99,134 | 2.7 | 795 | 7.8 |
| Trois-Rivieres (CMA) | 322 | 12 | 151,593 | 81,820 | 76,560 | 2.1 | 738 | 7.2 |
| Granby | 159 | 2 | 77,077 | 39,112 | 36,957 | 1.0 | 114 | 1.1 |
| Saint-Jean-sur-Richelieu | 184 | 3 | 92,394 | 43,555 | 42,305 | 1.2 | 130 | 1.3 |
| Montreal (CMA) | 7,297 | 58 | 3,824,225 | 1,802,903 | 1,713,615 | 47.3 | 2,057 | 20.2 |
| Gatineau (CMA) | 586 | 18 | 314,227 | 154,081 | 142,301 | 3.9 | 883 | 8.7 |
| Rouyn-Noranda/Val-d'Or | 142 | 5 | 73,398 | 39,569 | 37,029 | 1.0 | 207 | 2.0 |
| Saint-Georges | 71 | 3 | 34,642 | 17,816 | 16,622 | 0.5 | 196 | 1.9 |
| Other Urban | 2,333 | 36 | 1,149,527 | 598,274 | 562,191 | 15.5 | 1,828 | 17.9 |
| Non-Urban | 1,935 | 29 | 989,828 | 565,159 | 451,463 | 12.5 | 1,688 | 16.6 |
| Stratum type | | | | | | | | |
| Regular | 14,553 | 202 | 7,413,606 | 3,751,224 | 3,465,520 | 95.7 | 9,846 | 96.7 |
| High income | 743 | 8 | 417,715 | 162,782 | 155,433 | 4.3 | 339 | 3.3 |
| **Ontario** | **22,236** | **309** | **12,730,148** | **5,477,602** | **5,108,516** | **100.0** | **14,972** | **100.0** |
| EIER | | | | | | | | |
| 22 | 1,687 | 17 | 935,933 | 415,258 | 396,252 | 7.8 | 806 | 5.4 |
| 23 | 782 | 13 | 416,250 | 205,566 | 182,577 | 3.6 | 726 | 4.9 |
| 24 | 292 | 12 | 155,376 | 73,971 | 67,834 | 1.3 | 873 | 5.8 |
| 25 | 2,089 | 18 | 1,090,705 | 541,133 | 462,725 | 9.1 | 867 | 5.8 |
| 26 | 626 | 12 | 356,176 | 141,919 | 137,164 | 2.7 | 591 | 3.9 |
| 27 | 8,881 | 79 | 5,581,914 | 2,197,692 | 2,099,508 | 41.1 | 2,837 | 18.9 |
| 28 | 1,326 | 17 | 721,048 | 309,359 | 296,630 | 5.8 | 797 | 5.3 |
| 29 | 759 | 14 | 391,318 | 182,318 | 169,399 | 3.3 | 633 | 4.2 |
| 30 | 822 | 14 | 450,231 | 208,966 | 193,341 | 3.8 | 688 | 4.6 |
| 31 | 520 | 12 | 274,109 | 119,378 | 111,313 | 2.2 | 629 | 4.2 |
| 32 | 594 | 11 | 311,650 | 139,089 | 128,883 | 2.5 | 549 | 3.7 |
| 33 | 835 | 15 | 476,125 | 200,660 | 190,867 | 3.7 | 702 | 4.7 |
| 34 | 610 | 14 | 314,289 | 144,659 | 133,009 | 2.6 | 704 | 4.7 |
| 35 | 995 | 18 | 539,640 | 232,347 | 217,496 | 4.3 | 806 | 5.4 |
| 36 | 298 | 13 | 155,870 | 76,469 | 71,365 | 1.4 | 830 | 5.5 |
| 37 | 237 | 11 | 119,688 | 57,703 | 53,657 | 1.1 | 808 | 5.4 |
| 38 | 883 | 19 | 439,826 | 231,115 | 196,496 | 3.8 | 1,126 | 7.5 |
| ER | | | | | | | | |
| 3510 | 2,285 | 26 | 1,244,910 | 563,913 | 531,768 | 10.4 | 1,338 | 8.9 |
| 3515 | 817 | 18 | 445,204 | 221,807 | 195,081 | 3.8 | 1,168 | 7.8 |
| 3520 | 653 | 7 | 342,068 | 180,126 | 149,627 | 2.9 | 303 | 2.0 |
| 3530 | 9,396 | 90 | 5,877,168 | 2,316,567 | 2,213,911 | 43.3 | 3,397 | 22.7 |
| 3540 | 2,233 | 33 | 1,210,144 | 523,106 | 483,073 | 9.5 | 1,532 | 10.2 |
| 3550 | 2,564 | 42 | 1,363,189 | 602,342 | 569,131 | 11.1 | 2,018 | 13.5 |
| 3560 | 1,153 | 19 | 629,578 | 283,731 | 264,963 | 5.2 | 917 | 6.1 |
| 3570 | 1,188 | 25 | 616,906 | 279,804 | 258,160 | 5.1 | 1,233 | 8.2 |
| 3580 | 529 | 6 | 285,597 | 140,919 | 121,284 | 2.4 | 302 | 2.0 |
| 3590 | 1,019 | 28 | 520,573 | 269,509 | 235,513 | 4.6 | 1,682 | 11.2 |
| 3595 | 399 | 15 | 194,811 | 95,778 | 86,005 | 1.7 | 1,082 | 7.2 |

**Table B.4 (continued)**
**Characteristics of sample design by sub-provincial regions**

| Province | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| | number | | | | number | % | number | % |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Cornwall | 121 | 2 | 58,957 | 27,624 | 26,541 | 0.5 | 104 | 0.7 |
| Ottawa (CMA) | 1,663 | 17 | 921,824 | 409,551 | 390,751 | 7.6 | 795 | 5.3 |
| Kingston (CMA) | 293 | 12 | 156,144 | 74,755 | 68,162 | 1.3 | 877 | 5.9 |
| Peterborough (CMA) | 223 | 4 | 117,610 | 56,754 | 51,745 | 1.0 | 187 | 1.2 |
| Oshawa (CMA) | 626 | 12 | 356,176 | 141,919 | 137,164 | 2.7 | 591 | 3.9 |
| Toronto (CMA) | 8,881 | 79 | 5,581,914 | 2,197,692 | 2,099,508 | 41.1 | 2,837 | 18.9 |
| Hamilton (CMA) | 1,326 | 17 | 721,048 | 309,359 | 296,630 | 5.8 | 797 | 5.3 |
| St-Catharines/Niagara (CMA) | 759 | 14 | 391,318 | 182,318 | 169,399 | 3.3 | 633 | 4.2 |
| Kitchener/Camb./Wat. (CMA) | 835 | 15 | 476,125 | 200,660 | 190,867 | 3.7 | 702 | 4.7 |
| Brantford (CMA) | 246 | 6 | 129,288 | 55,589 | 53,131 | 1.0 | 336 | 2.2 |
| Norfolk | 123 | 3 | 62,822 | 29,088 | 26,572 | 0.5 | 134 | 0.9 |
| Guelph (CMA) | 260 | 6 | 140,802 | 62,481 | 58,021 | 1.1 | 270 | 1.8 |
| London (CMA) | 869 | 15 | 474,237 | 218,577 | 202,713 | 4.0 | 721 | 4.8 |
| Chatham-Kent | 210 | 5 | 103,671 | 48,265 | 44,715 | 0.9 | 239 | 1.6 |
| Windsor (CMA) | 605 | 12 | 319,247 | 142,690 | 131,879 | 2.6 | 562 | 3.8 |
| Sarnia | 178 | 4 | 88,915 | 41,944 | 39,716 | 0.8 | 213 | 1.4 |
| Barrie (CMA) | 365 | 4 | 187,013 | 76,350 | 71,872 | 1.4 | 193 | 1.3 |
| North Bay | 126 | 3 | 64,163 | 31,204 | 29,118 | 0.6 | 149 | 1.0 |
| Greater Sudbury (CMA) | 306 | 13 | 160,274 | 78,566 | 73,344 | 1.4 | 853 | 5.7 |
| Sault Ste-Marie | 163 | 4 | 78,693 | 37,424 | 35,426 | 0.7 | 181 | 1.2 |
| Thunder Bay (CMA) | 239 | 11 | 120,736 | 58,299 | 54,091 | 1.1 | 815 | 5.4 |
| Leamington | 84 | 2 | 49,765 | **20,247** | **18,928** | **0.4** | **101** | **0.7** |
| Timmins | **76** | **2** | **43,165** | 20,284 | 19,217 | 0.4 | 98 | 0.7 |
| Other Urban | 2,079 | 29 | 1,088,446 | 527,628 | 483,428 | 9.5 | 1,476 | 9.9 |
| Non-Urban | 1,580 | 19 | 837,795 | 428,334 | 335,578 | 6.6 | 1,107 | 7.4 |
| Stratum type | | | | | | | | |
| Regular | 20,398 | 283 | 11,507,031 | 5,010,408 | 4,663,465 | 91.3 | 13,755 | 91.9 |
| High income | 1,462 | 17 | 895,718 | 302,548 | 294,069 | 5.8 | 738 | 4.9 |
| Large-cluster | 153 | 4 | 223,170 | 111,764 | 105,612 | 2.1 | 143 | 1.0 |
| Remote | 223 | 5 | 104,229 | 52,882 | 45,370 | 0.9 | 336 | 2.2 |
| **Manitoba** | **2,098** | **79** | **1,137,354** | **504,897** | **467,146** | **100.0** | **4,206** | **100.0** |
| EIER | | | | | | | | |
| 39 | 1,316 | 48 | 723,733 | 318,220 | 303,303 | 64.9 | 2,400 | 57.1 |
| 40 | 626 | 22 | 336,826 | 143,707 | 132,184 | 28.3 | 1,092 | 26.0 |
| 41 | 156 | 9 | 76,795 | 42,970 | 31,659 | 6.8 | 714 | 17.0 |
| ER | | | | | | | | |
| 4610 | 178 | 6 | 101,814 | 42,443 | 37,108 | 7.9 | 443 | 10.5 |
| 4620 | 103 | 4 | 60,285 | 23,032 | 22,102 | 4.7 | 178 | 4.2 |
| 4630 | 208 | 7 | 105,934 | 50,037 | 44,998 | 9.6 | 329 | 7.8 |
| 4640 | 83 | 3 | 44,427 | 17,168 | 16,090 | 3.4 | 129 | 3.1 |
| 4650 | 1,220 | 46 | 666,760 | 296,287 | 283,024 | 60.6 | 2,242 | 53.3 |
| 4660 | 152 | 6 | 86,657 | 41,411 | 33,835 | 7.2 | 407 | 9.7 |
| 4670 | 84 | 3 | 38,953 | 19,575 | 17,238 | 3.7 | 149 | 3.5 |
| 4680 | 70 | 5 | 32,524 | 14,944 | 12,751 | 2.7 | 329 | 7.8 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Winnipeg (CMA) | 1,327 | 48 | 730,013 | 320,336 | 305,366 | 65.4 | 2,416 | 57.4 |
| Brandon | 101 | 4 | 53,229 | 24,342 | 23,241 | 5.0 | 170 | 4.0 |
| Thompson | 26 | 2 | 12,829 | 5,486 | 4,821 | 1.0 | 124 | 3.0 |
| Other Urban | 249 | 12 | 135,516 | 61,082 | 57,869 | 12.4 | 625 | 14.9 |
| Non-Urban | 395 | 13 | 205,767 | 93,651 | 75,849 | 16.2 | 870 | 20.7 |
| Stratum type | | | | | | | | |
| Regular | 1,847 | 68 | 998,748 | 452,489 | 418,266 | 89.5 | 3,594 | 85.4 |
| High income | 126 | 4 | 73,108 | 24,695 | 24,268 | 5.2 | 192 | 4.6 |
| Immigrant | 55 | 2 | 32,974 | 12,769 | 11,861 | 2.5 | 91 | 2.2 |
| Remote | 70 | 5 | 32,524 | 14,944 | 12,751 | 2.7 | 329 | 7.8 |

**Table B.4 (continued)**
**Characteristics of sample design by sub-provincial regions**

| Province | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| | number | | | | number | % | number | % |
| **Saskatchewan** | **1,881** | **67** | **946,824** | **436,428** | **400,608** | **100.0** | **4,122** | **100.0** |
| EIER | | | | | | | | |
| 42 | 419 | 15 | 210,410 | 96,261 | 91,367 | 22.8 | 868 | 21.1 |
| 43 | 480 | 18 | 260,249 | 116,976 | 110,387 | 27.6 | 1,056 | 25.6 |
| 44 | 607 | 18 | 285,062 | 138,379 | 122,755 | 30.6 | 1,038 | 25.2 |
| 45 | 375 | 16 | 191,103 | 84,812 | 76,099 | 19.0 | 1,160 | 28.1 |
| ER | | | | | | | | |
| 4710 | 561 | 19 | 280,215 | 129,336 | 120,764 | 30.1 | 1,099 | 26.7 |
| 4720 | 213 | 7 | 95,908 | 46,827 | 41,711 | 10.4 | 355 | 8.6 |
| 4730 | 577 | 20 | 306,992 | 138,486 | 129,710 | 32.4 | 1,229 | 29.8 |
| 4740 | 163 | 5 | 76,671 | 38,633 | 33,931 | 8.5 | 303 | 7.4 |
| 4750 | 347 | 15 | 174,638 | 78,565 | 70,392 | 17.6 | 1,073 | 26.0 |
| 4760 | 20 | 1 | 12,400 | 4,581 | 4,100 | 1.0 | 62 | 1.5 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Regina (CMA) | 419 | 15 | 210,410 | 96,261 | 91,367 | 22.8 | 868 | 21.1 |
| Saskatoon (CMA) | 480 | 18 | 260,249 | 116,976 | 110,387 | 27.6 | 1,056 | 25.6 |
| Moose Jaw | 74 | 3 | 34,512 | 16,832 | 15,895 | 4.0 | 135 | 3.3 |
| Prince Albert | 78 | 3 | 42,507 | 17,584 | 16,323 | 4.1 | 249 | 6.0 |
| Other Urban | 360 | 13 | 179,748 | 83,221 | 78,000 | 19.5 | 841 | 20.4 |
| Non-Urban | 470 | 15 | 219,398 | 105,554 | 88,636 | 22.1 | 973 | 23.6 |
| Stratum type | | | | | | | | |
| Regular | 1,728 | 60 | 857,680 | 402,423 | 368,604 | 92.0 | 3,793 | 92.0 |
| Aboriginal | 66 | 3 | 34,197 | 14,954 | 13,743 | 3.4 | 131 | 3.2 |
| High income | 67 | 3 | 42,547 | 14,470 | 14,161 | 3.5 | 135 | 3.3 |
| Remote | 20 | 1 | 12,400 | 4,581 | 4,100 | 1.0 | 62 | 1.5 |
| **Alberta** | **6,379** | **88** | **3,557,377** | **1,606,911** | **1,487,737** | **100.0** | **4,500** | **100.0** |
| EIER | | | | | | | | |
| 46 | 2,096 | 22 | 1,213,088 | 531,335 | 503,607 | 33.9 | 1,023 | 22.7 |
| 47 | 2,039 | 20 | 1,156,330 | 527,074 | 492,333 | 33.1 | 966 | 21.5 |
| 48 | 402 | 19 | 219,533 | 101,180 | 85,891 | 5.8 | 1,219 | 27.1 |
| 49 | 1,842 | 27 | 968,426 | 447,322 | 405,906 | 27.3 | 1,292 | 28.7 |
| ER | | | | | | | | |
| 4810 | 486 | 10 | 263,065 | 120,431 | 110,508 | 7.4 | 458 | 10.2 |
| 4820 | 370 | 5 | 188,248 | 85,872 | 78,273 | 5.3 | 256 | 5.7 |
| 4830 | 2,267 | 23 | 1,308,684 | 573,157 | 542,207 | 36.4 | 1,099 | 24.4 |
| 4840 | 150 | 1 | 77,699 | 40,338 | 33,612 | 2.3 | 83 | 1.8 |
| 4850 | 369 | 5 | 184,575 | 86,584 | 78,714 | 5.3 | 244 | 5.4 |
| 4860 | 2,106 | 21 | 1,195,404 | 545,189 | 508,837 | 34.2 | 1,017 | 22.6 |
| 4870 | 420 | 10 | 219,245 | 100,116 | 88,673 | 6.0 | 646 | 14.4 |
| 4880 | 211 | 11 | 120,457 | 55,224 | 46,913 | 3.2 | 696 | 15.5 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Medicine Hat | 139 | 2 | 71,178 | 34,740 | 32,133 | 2.2 | 93 | 2.1 |
| Lethbridge | 202 | 6 | 106,372 | 51,205 | 46,935 | 3.2 | 273 | 6.1 |
| Calgary (CMA) | 2,096 | 22 | 1,213,088 | 531,335 | 503,607 | 33.9 | 1,023 | 22.7 |
| Red Deer | 183 | 2 | 90,564 | 42,740 | 39,968 | 2.7 | 124 | 2.7 |
| Edmonton (CMA) | 2,039 | 20 | 1,156,330 | 527,074 | 492,333 | 33.1 | 966 | 21.5 |
| Wood Buffalo | 101 | 5 | 62,239 | 28,574 | 23,387 | 1.6 | 347 | 7.7 |
| Other Urban | 901 | 16 | 478,801 | 228,491 | 207,711 | 14.0 | 882 | 19.6 |
| Non-Urban | 718 | 15 | 378,805 | 162,752 | 141,663 | 9.5 | 792 | 17.6 |
| Stratum type | | | | | | | | |
| Regular | 5,594 | 77 | 3,101,778 | 1,418,114 | 1,311,882 | 88.2 | 4,008 | 89.1 |
| Aboriginal | 551 | 8 | 302,531 | 134,499 | 123,679 | 8.3 | 344 | 7.6 |
| High income | 213 | 2 | 140,572 | 49,991 | 48,329 | 3.2 | 96 | 2.1 |
| Remote | 21 | 1 | 12,496 | 4,307 | 3,847 | 0.3 | 52 | 1.1 |

**Table B.4 (end)**
**Characteristics of sample design by sub-provincial regions**

| Province | PSUs | Strata | Population | Dwellings | Households | | Sampled households | |
|---|---|---|---|---|---|---|---|---|
| | number | | | | number | % | number | % |
| **British Columbia** | **7,909** | **118** | **4,267,875** | **2,048,300** | **1,881,831** | **100.0** | **5,427** | **100.0** |
| EIER | | | | | | | | |
| 50 | 1,238 | 17 | 634,074 | 337,928 | 299,073 | 15.9 | 790 | 14.6 |
| 51 | 283 | 11 | 169,924 | 69,132 | 64,213 | 3.4 | 663 | 12.2 |
| 52 | 4,069 | 41 | 2,301,575 | 1,042,496 | 977,381 | 51.9 | 1,578 | 29.1 |
| 53 | 676 | 17 | 338,615 | 179,073 | 165,293 | 8.8 | 829 | 15.3 |
| 54 | 1,069 | 18 | 535,082 | 279,895 | 251,482 | 13.4 | 842 | 15.5 |
| 55 | 574 | 14 | 288,605 | 139,776 | 124,389 | 6.6 | 725 | 13.4 |
| ER | | | | | | | | |
| 5910 | 1,456 | 29 | 722,820 | 381,455 | 350,067 | 18.6 | 1,412 | 26.0 |
| 5920 | 4,659 | 58 | 2,630,450 | 1,193,330 | 1,111,808 | 59.1 | 2,525 | 46.5 |
| 5930 | 953 | 12 | 492,573 | 259,231 | 231,736 | 12.3 | 557 | 10.3 |
| 5940 | 285 | 5 | 141,501 | 78,697 | 67,337 | 3.6 | 233 | 4.3 |
| 5950 | 286 | 6 | 145,526 | 72,346 | 64,669 | 3.4 | 262 | 4.8 |
| 5960 | 86 | 2 | 40,732 | 20,105 | 17,610 | 0.9 | 123 | 2.3 |
| 5970 | 70 | 2 | 34,835 | 15,565 | 13,866 | 0.7 | 99 | 1.8 |
| 5980 | 114 | 4 | 59,438 | 27,571 | 24,738 | 1.3 | 216 | 4.0 |
| CMA/CA-Design Type Indicator | | | | | | | | |
| Kelowna (CMA) | 308 | 5 | 171,018 | 88,648 | 79,290 | 4.2 | 249 | 4.6 |
| Chilliwack | 165 | 3 | 88,128 | 40,789 | 38,117 | 2.0 | 154 | 2.8 |
| Abbotsford - Mission (CMA) | 283 | 11 | 169,924 | 69,132 | 64,213 | 3.4 | 663 | 12.2 |
| Vancouver (CMA) | 4,069 | 41 | 2,301,575 | 1,042,496 | 977,381 | 51.9 | 1,578 | 29.1 |
| Victoria (CMA) | 676 | 17 | 338,615 | 179,073 | 165,293 | 8.8 | 829 | 15.3 |
| Nanaimo | 206 | 3 | 97,153 | 49,989 | 46,729 | 2.5 | 144 | 2.7 |
| Prince George | 160 | 3 | 83,764 | 40,156 | 36,700 | 2.0 | 149 | 2.7 |
| Fort St. John | 50 | 2 | 26,380 | 12,041 | 11,087 | 0.6 | 97 | 1.8 |
| Other Urban | 1,487 | 23 | 741,046 | 381,504 | 347,422 | 18.5 | 1,144 | 21.1 |
| Non-Urban | 505 | 9 | 250,272 | 144,472 | 115,599 | 6.1 | 420 | 7.7 |
| Stratum type | | | | | | | | |
| Regular | 6,685 | 99 | 3,606,814 | 1,751,931 | 1,611,838 | 85.7 | 4,558 | 84.0 |
| Aboriginal | 610 | 9 | 327,646 | 156,267 | 143,433 | 7.6 | 379 | 7.0 |
| High income | 393 | 5 | 225,316 | 85,945 | 81,194 | 4.3 | 206 | 3.8 |
| Remote | 221 | 5 | 108,099 | 54,157 | 45,366 | 2.4 | 284 | 5.2 |
| **Canada (provinces only)** | **60,136** | **1,153** | **32,751,627** | **15,130,549** | **13,981,923** | **100.0** | **52,617** | **100.0** |
| Stratum type | | | | | | | | |
| Regular | 54,908 | 930 | 29,570,025 | 13,838,102 | 12,775,134 | 91.4 | 47,201 | 89.7 |
| Aboriginal | 1,227 | 20 | 664,374 | 305,720 | 280,855 | 2.0 | 854 | 1.6 |
| High income | 3,100 | 42 | 1,851,912 | 661,336 | 637,625 | 4.6 | 1,844 | 3.5 |
| Immigrant | 55 | 2 | 32,974 | 12,769 | 11,861 | 0.1 | 91 | 0.2 |
| Large-cluster | 153 | 4 | 223,170 | 111,764 | 105,612 | 0.8 | 143 | 0.3 |
| One-stage | 138 | 138 | 139,424 | 69,987 | 59,402 | 0.4 | 1,421 | 2.7 |
| Remote | 555 | 17 | 269,748 | 130,871 | 111,434 | 0.8 | 1,063 | 2.0 |

**Note:** See Appendix A.2 for abbreviations

### List of variables used for the optimal stratification

The list of variables used for the optimal stratification is identical to that used in the last redesign.

The choice of stratification variables was adapted to each region for which optimal stratification was used. For each PSU in the region, the variables below were obtained from Census 2011 or NHS 2011 data. If a variable represented less than 2% of the total population, it was dropped. For categories such as services, if a sub-category, such as financial services, had too few employed persons, then the global variable was used instead. A category was considered significant if it represented more than 2% of the population.

Number of persons employed in the following sectors:

- Agriculture
- Forestry and fishing
- Mining
- Manufacturing - consumables
- Manufacturing - rubber, plastics, leather
- Manufacturing - textiles and clothing
- Manufacturing - furniture, pulp and paper, printing, wood
- Manufacturing - metals and minerals
- Manufacturing - petrochemical, chemical
- Construction
- Transportation
- Services - trade
- Services - financial
- Services - personal/business
- Services - government

Total employed

Total income

Population aged 15+

Population aged 15 to 24

Population aged 55+

Number of one-person households

Number of two-person households

Number of owned dwellings

Total gross rent

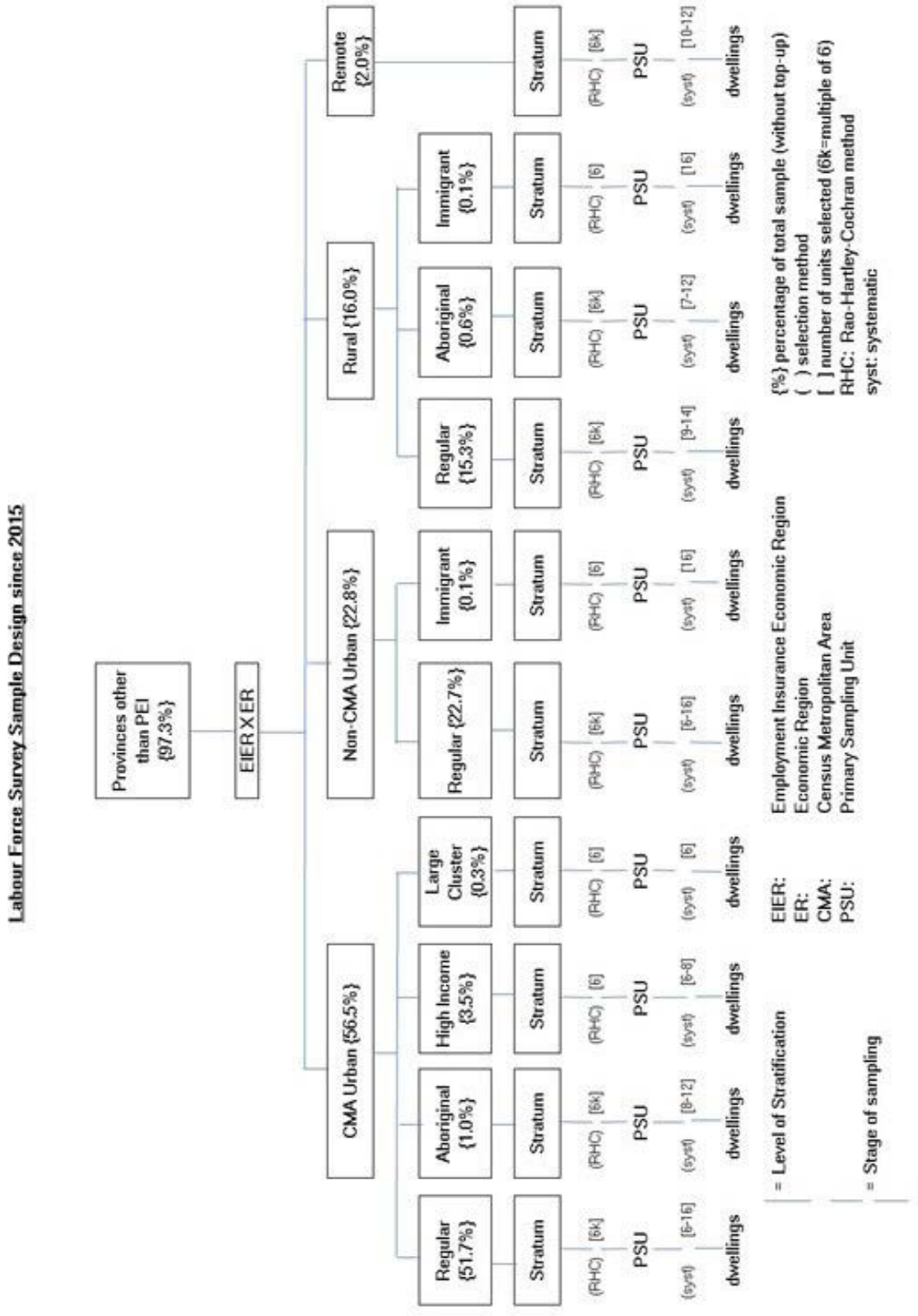
Population with high school education

Mother tongue English

Mother tongue French

Mother tongue other than English/French

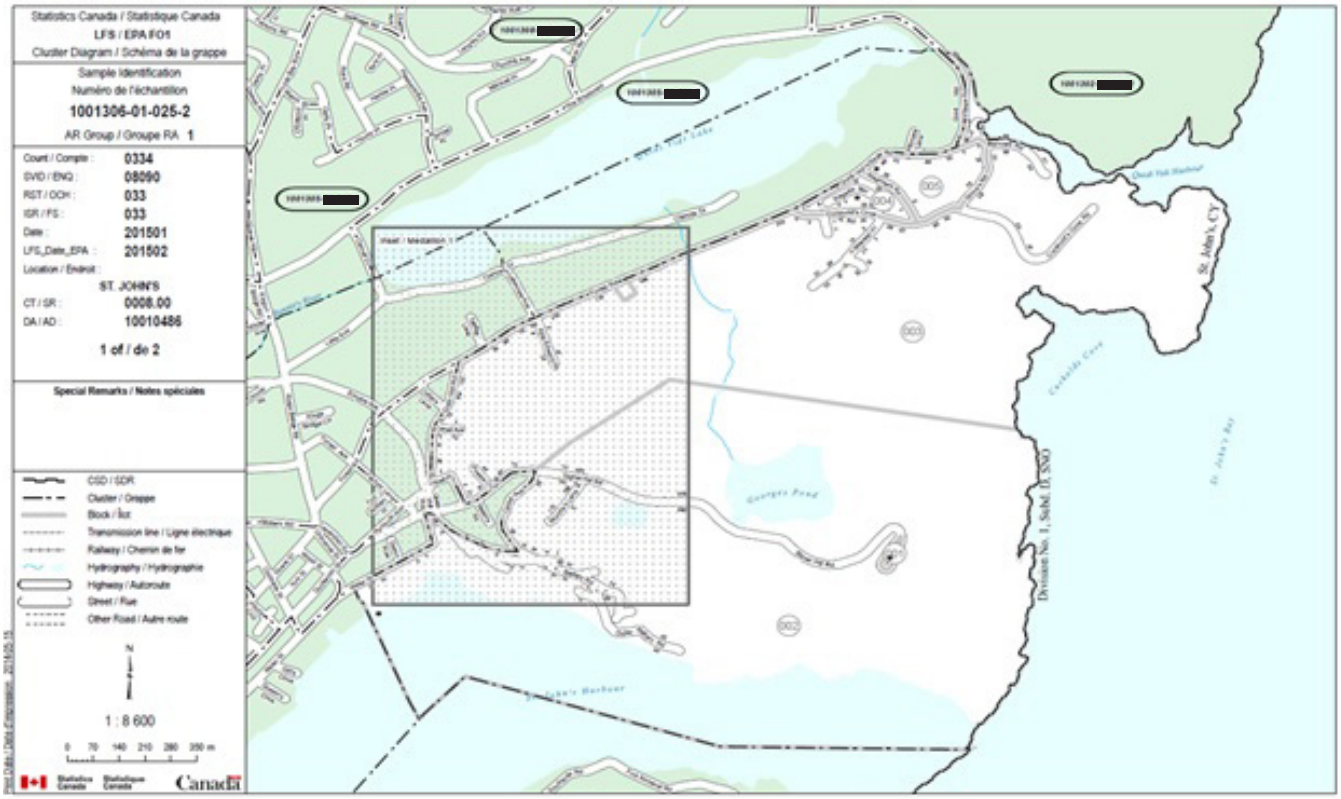
# Appendix C Labour Force Survey Sample Design



Labour Force Survey Sample Design since 2015
Prince Edward Island {2.7%}
EIER X ER
Super Stratum
Stratum
(syst) [10]
dwellings
= Level of Stratification
= Stage of sampling
EIER: Employment Insurance Economic Region
ER: Economic Region
CMA: Census Metropolitan Area
PSU: Primary Sampling Unit
{%} percentage of total sample (without top-up)
( ) selection method
[ ] number of units selected (6k=multiple of 6)
RHC: Rao-Hartley-Cochran method
syst: systematic

# Appendix D PSU maps (F01 cluster diagrams)

PSU maps are generated using Generalized Mapping System (GMS) software in place since 2009. Geography Division has also developed a cluster viewer application. This is an electronic map of Canada with a number of features that allow the user to interactively zoom to specific PSUs, change the range or scale of the map, as well as to create .pdfs files that are similar to those produced by the GMS. A table of PSUs along with the relevant information for the map legend is one of the key GMS inputs. All maps are carefully checked before shipment to the Regional Offices (RO). Map sizes and inset boundaries are automatically generated but may not be appropriate in all cases. Clerical verification of the PSU maps is done to improve the final product. The GMS has tools that allow adjustment of the map size, as well as deletion or creation of insets. Below is an example of a PSU map. The legend is explained on the next page, followed by the inset map (portrait orientation).



Artificial IDs, dates, and counts have been used in this example.

**Sample Identification** is the PSU ID, fake for this example, (stratum 1001306; design type 01; PSU 030; rotation group 2).

**AR Group:** Classifies the PSU based on their membership in the mail-out for Census and expected AR listing quality (0 = mail-out, high quality; 1 = non-mail-out, high quality, 2 = non-mail-out, low quality)

**Count:** Design count of the number of households in the PSU (334)

**SVID:** Survey ID of the first survey to use the PSU (8090 is the CCHS, 10440 is the LFS, etc.)

**RST:** First random start used for dwelling selection

**ISR:** PSU level inverse sampling rate used for dwelling selection

**Date:** Date of first use of this PSU by any survey

**LFS_Date_EPA:** Date of first LFS use. If SVID is 10440 = LFS, then LFS_Date_EPA is set to 0 by convention.

**Location:** A Canada Post municipality. Usually, the most frequent one appearing the prospective dwelling list.

**CT:** Census Tract code from Census base at the time of PSU introduction. This will be Census 2011 initially and Census 2016 later on.

**DA:** Dissemination Area code from Census base at the time of PSU introduction. This will be Census 2011 initially and Census 2016 later on.

**N1 of / de N2:** Map number and total number of maps for this PSU (1 of 2). Two PSU maps means there is one main map (number 1) and one inset (number 2).

**Inset:** For inset maps, the inset number. Inset does not appear in legend of main maps.

**Special Remarks:** Extra pertinent information about the PSU – usually blank.

**Symbol legend** shows how geography features are displayed.

**North symbol** always points up.

**Scale** of the map.

**Print Date:** On landscape maps, Print Date appears in the lower left, rotated and outside of legend. On portrait maps, Print Date appears in the bottom centre, below the legend.

Area outside PSU is green; area inside is yellow.

Blocks are numbered and circled.

Starting point is an asterisk.

Inset rectangles are dotted gray.

Neighbouring PSUs are labeled. Specific PSU blanked out in the example.

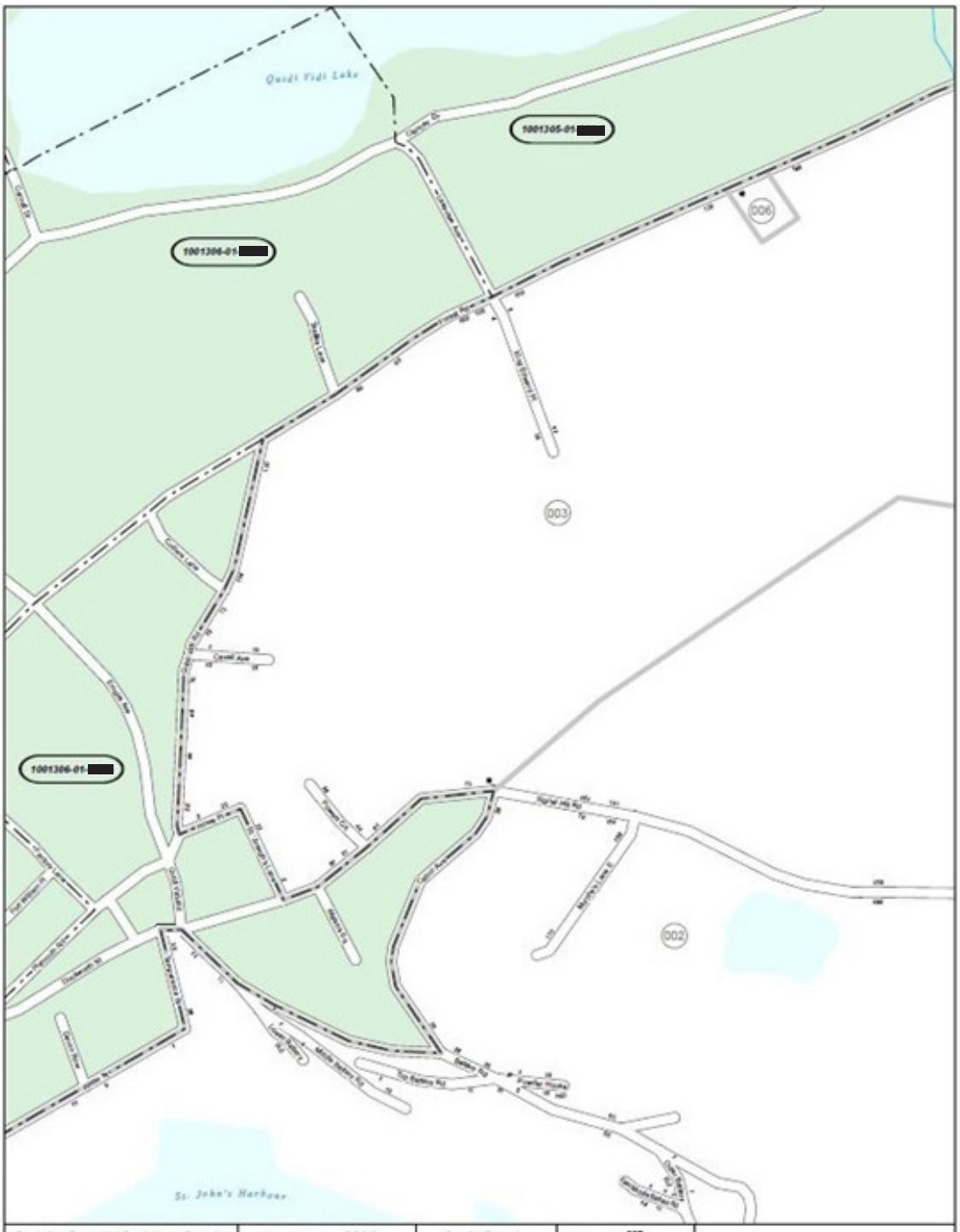
Address ranges are shown along roads.

Street labels are more legible on the actual maps than on these images.

Maps can be oriented either landscape (legend at left) or portrait (legend at bottom).

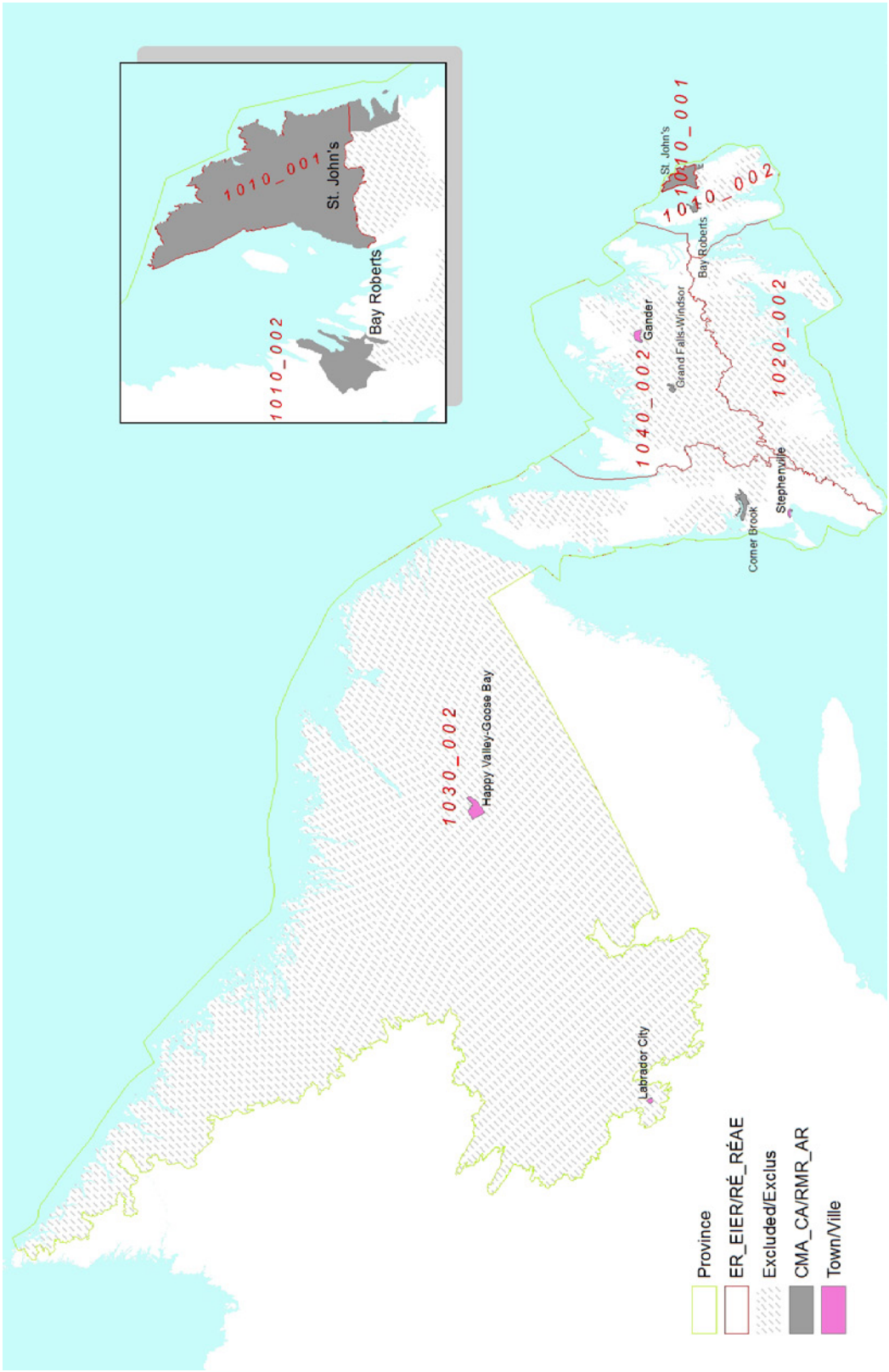
The possible main map sizes are 11” x 17”, 17” x 22”, or 22” x 34”.

All insets are 11” x 17”.

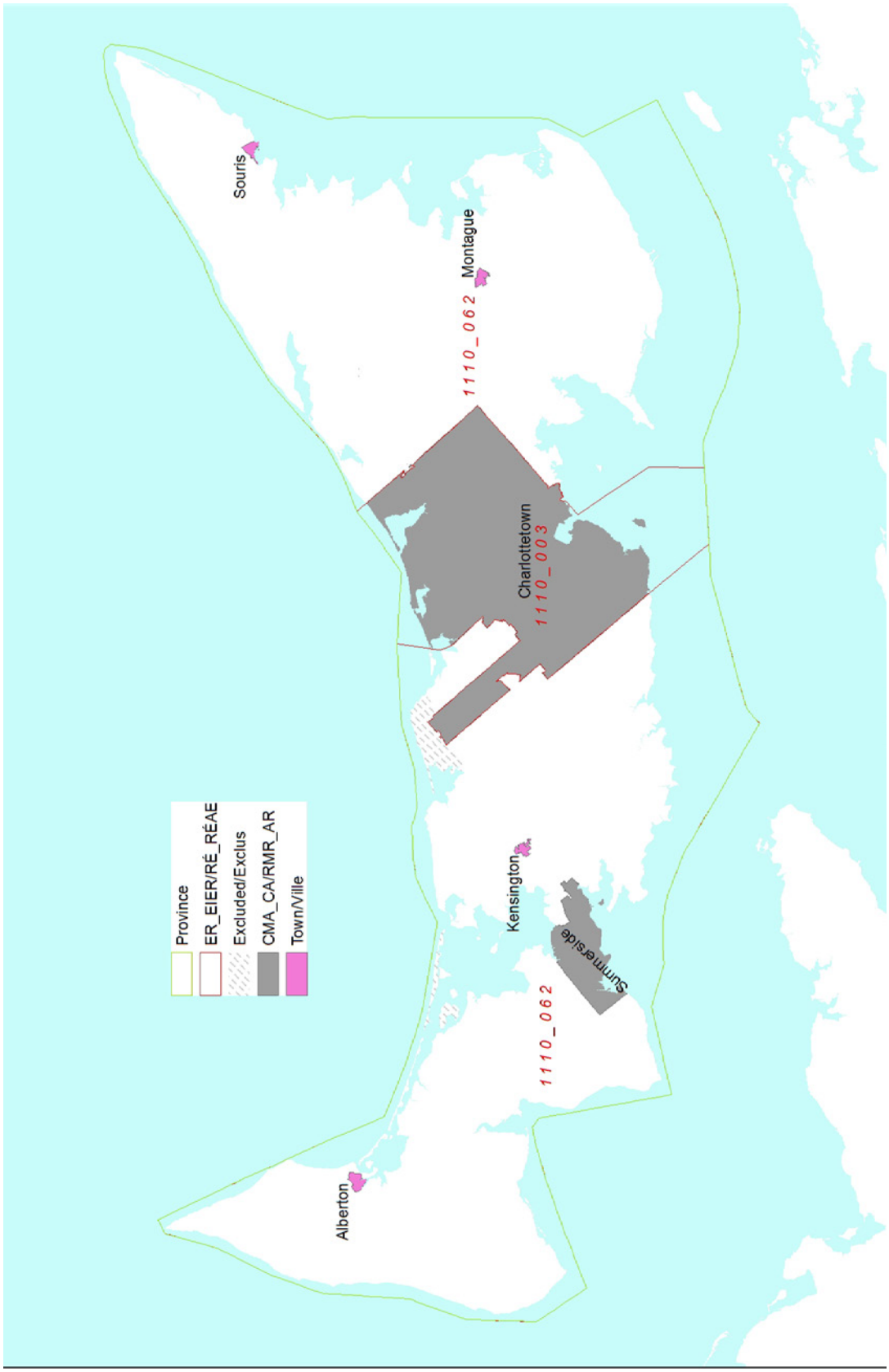

Quidi Vidi Lake
1001305-01
1001306-01
1001306-01
206
203
202
St. John's Harbour
Statistics Canada / Statistique Canada
LFS / EPA FO1
Cluster Diagram / Schéma de la grappe
Sample Identification
Numéro de l'échantillon
1001306-01-025-2
AR Group / Groupe RA 1
2 of / de 2
Inset / Médaillon 1
Count / Compte : 0334
SVID / ENQ : 08090
RST / OCH : 033
ISR / FS : 033
Date : 201501
LFS_Date_EPA : 201502
Location / Endroit :
ST. JOHN'S
CT / SR : 0008.00
DA / AD : 10010486
Special Remarks
Notes spéciales
CSD
SDR
Cluster
Grappe
Block
Ilot
Transmission line
Ligne électrique
Railway
Chemin de fer
Hydrography
Hydrographie
Highway
Autoroute
Street
Rue
Other Road
Autre route
N
1 : 3 200
0 25 50 75 100 125 m
Print Date / Date d'impression 2014-05-15
Canada

# Appendix E Provincial maps

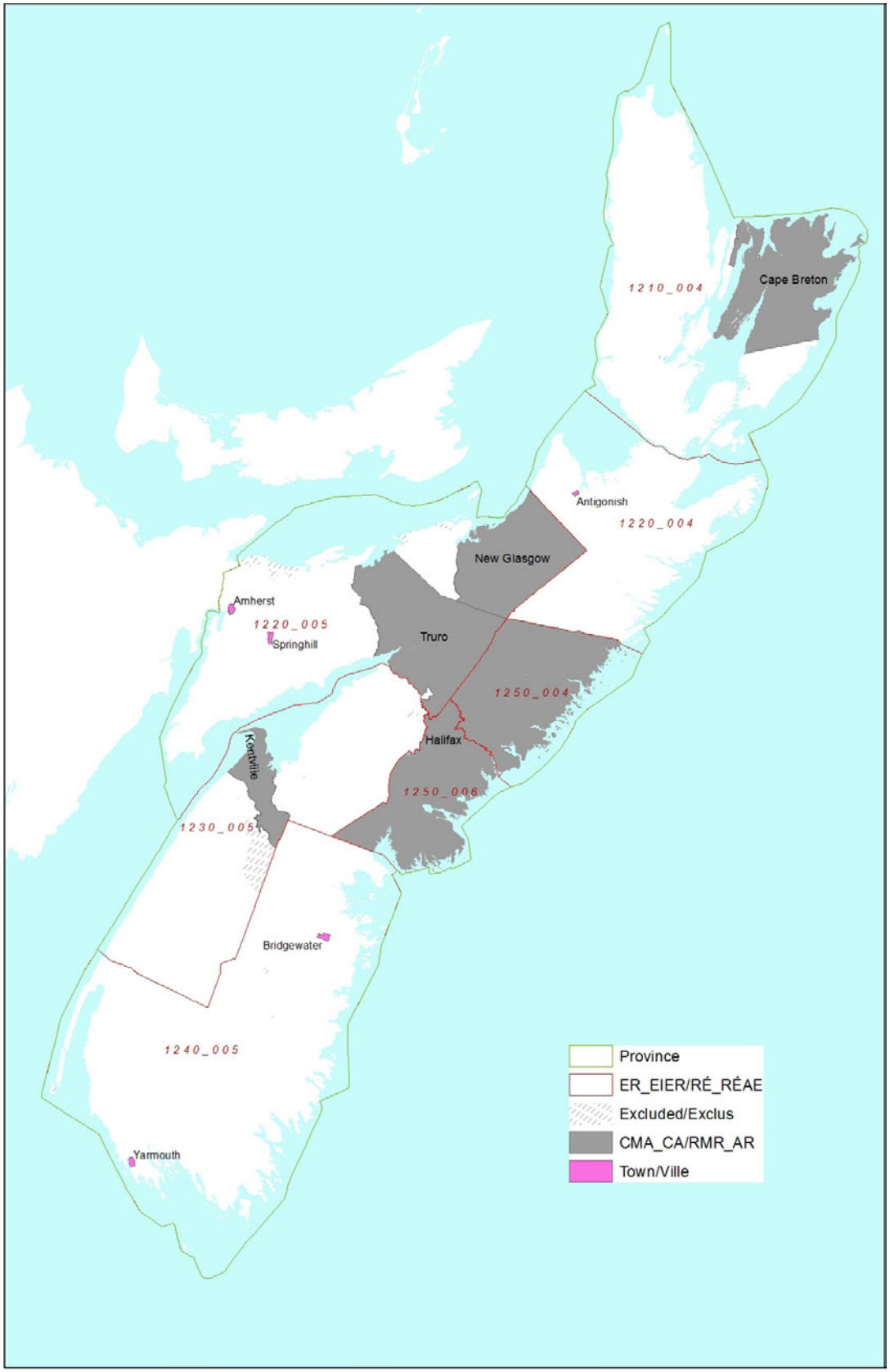
**Map 1 – Newfoundland and Labrador**
**CMACA, EI-EIER Intersection Boundaries**



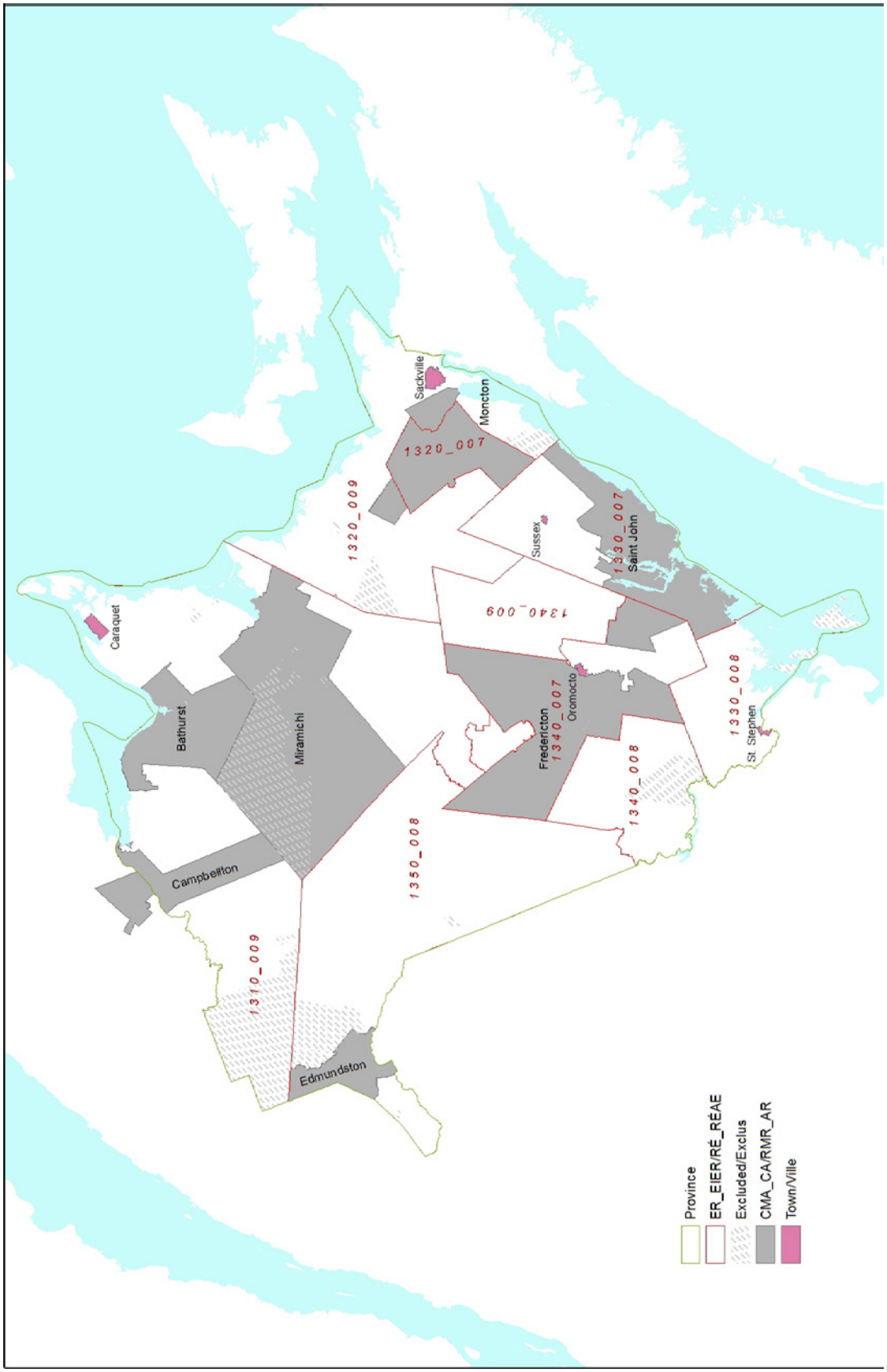
**Map 2 – Prince Edward Island**
**CMACA, EI-EIER Intersection Boundaries**



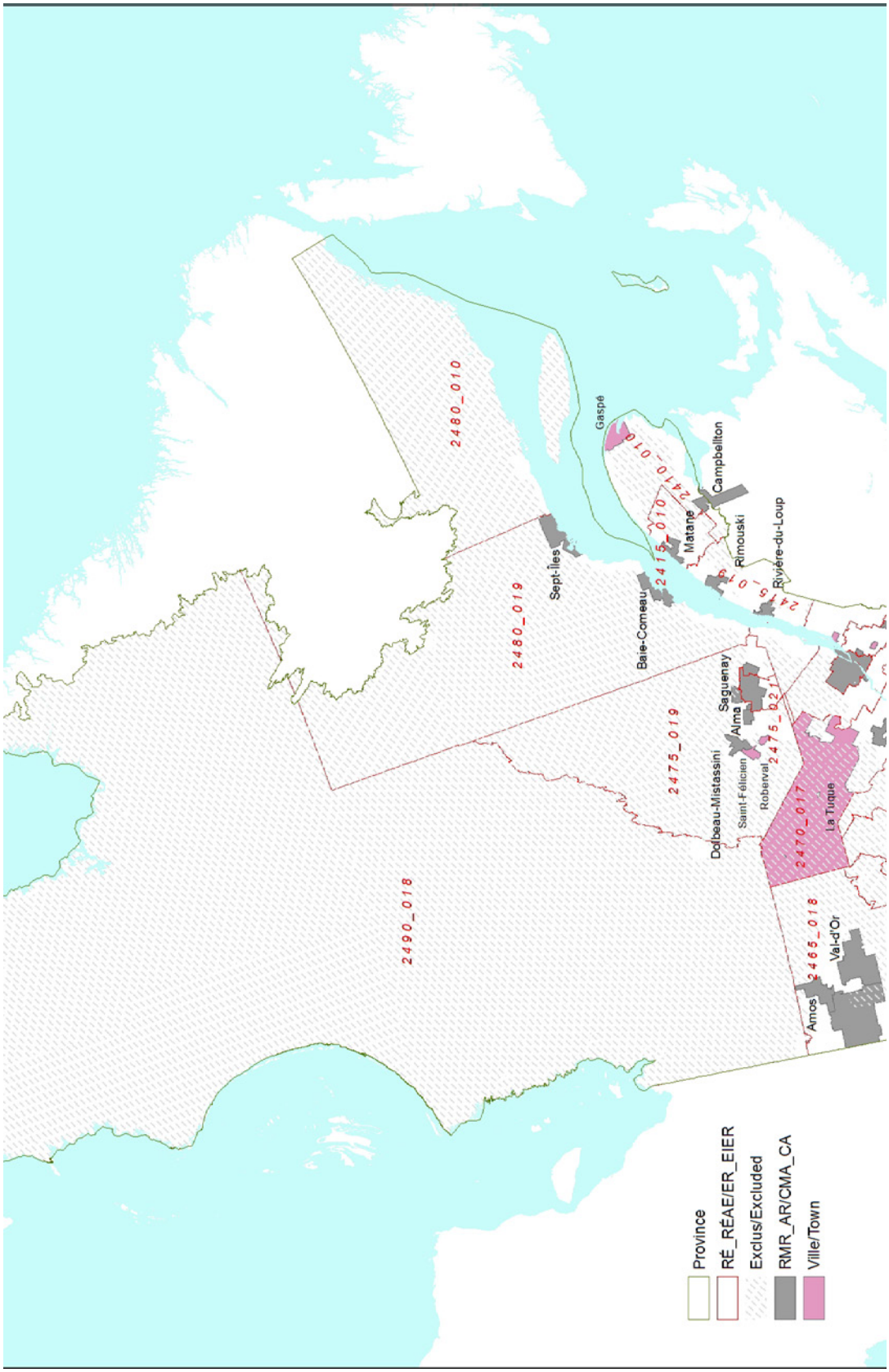
**Map 3 – Nova Scotia**
**CMACA, EI-EIER Intersection Boundaries**



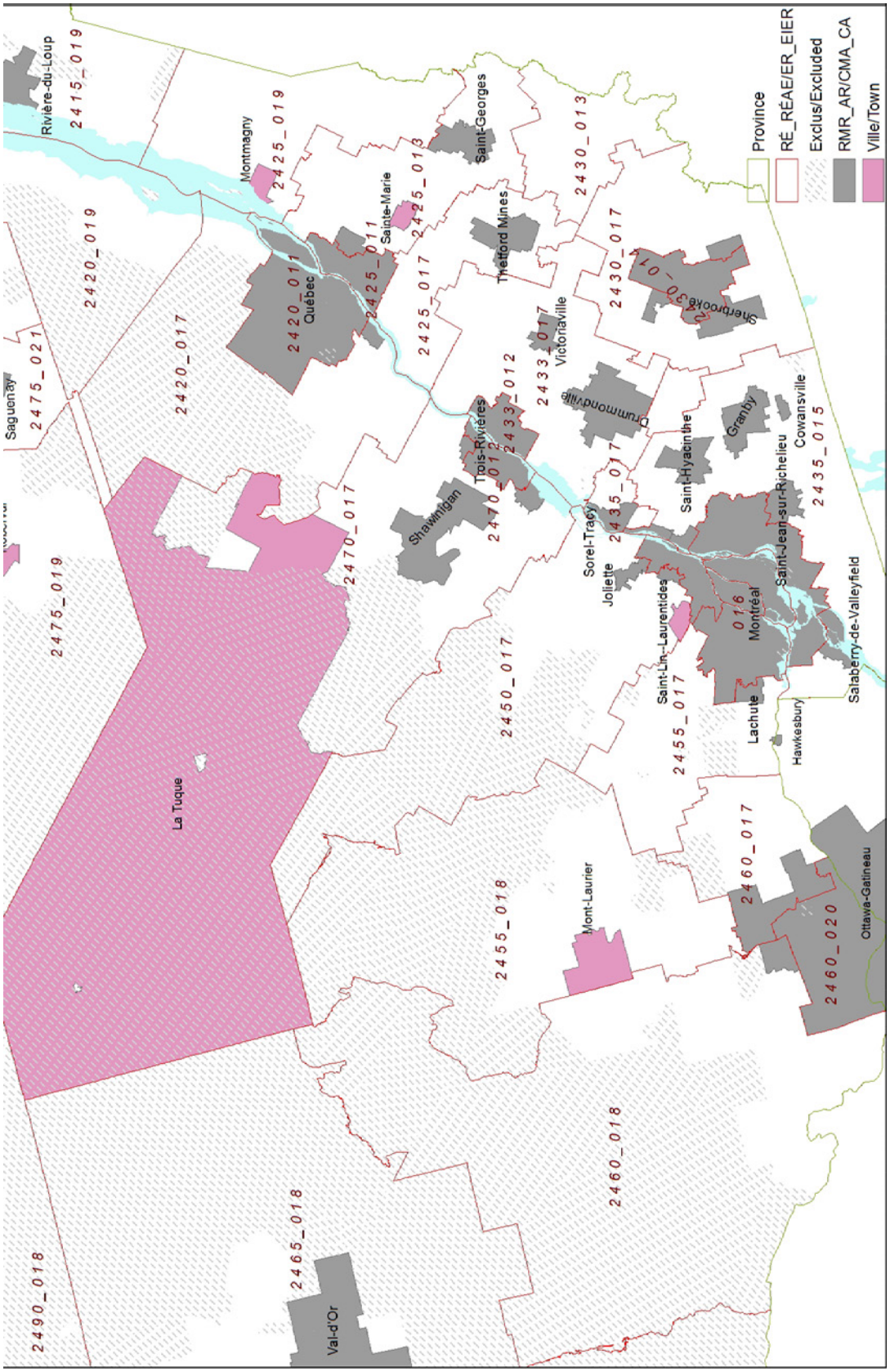
**Map 4 – New Brunswick**
**CMACA, EI-EIER Intersection Boundaries**



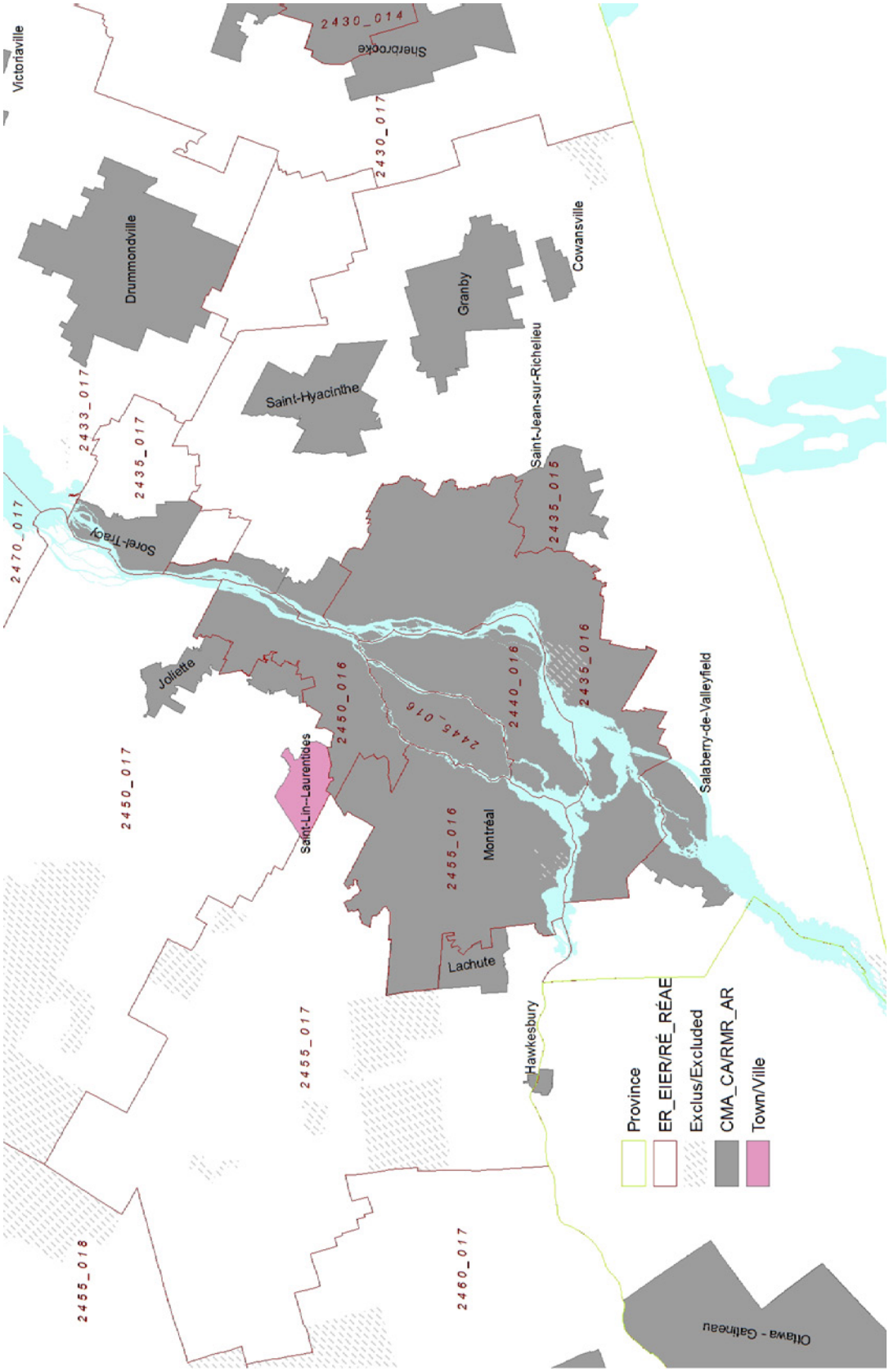
**Map 5a – Quebec North**
**CMACA, EI-EIER Intersection Boundaries**



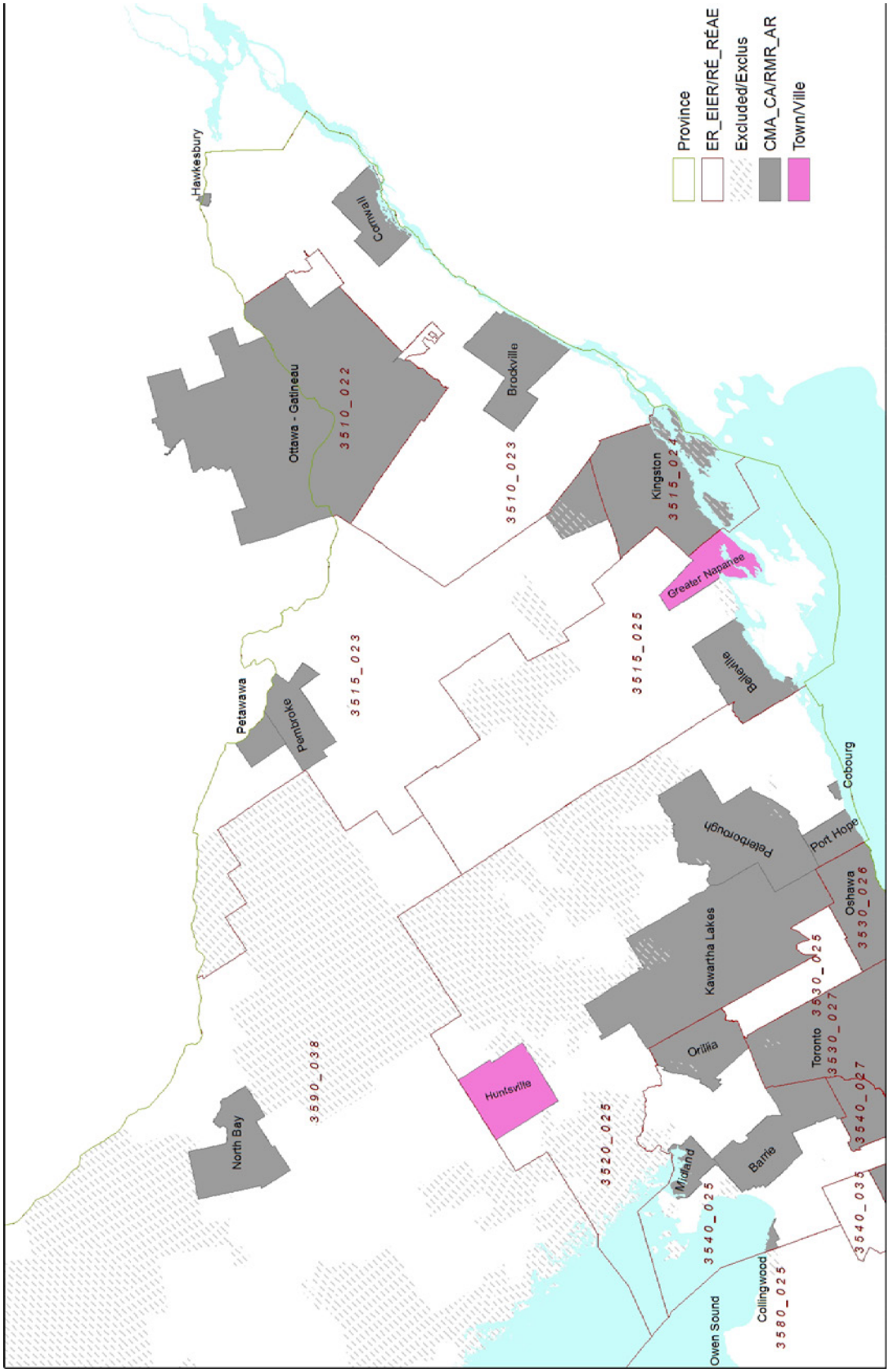
**Map 5b – Quebec South**
**CMACA, EI-EIER Intersection Boundaries**



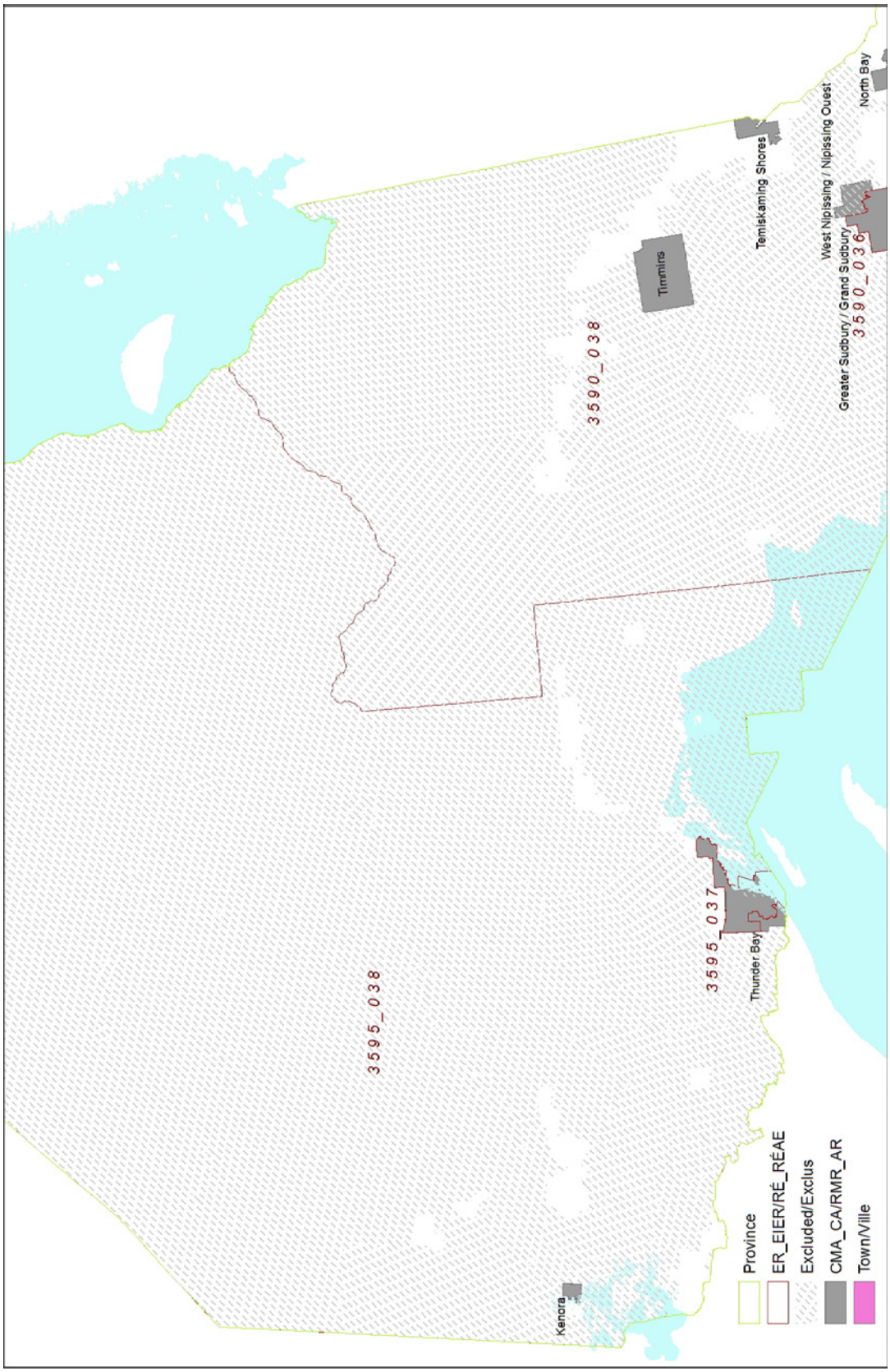
**Map 5c – Montreal**
**CMACA, EI-EIER Intersection Boundaries**



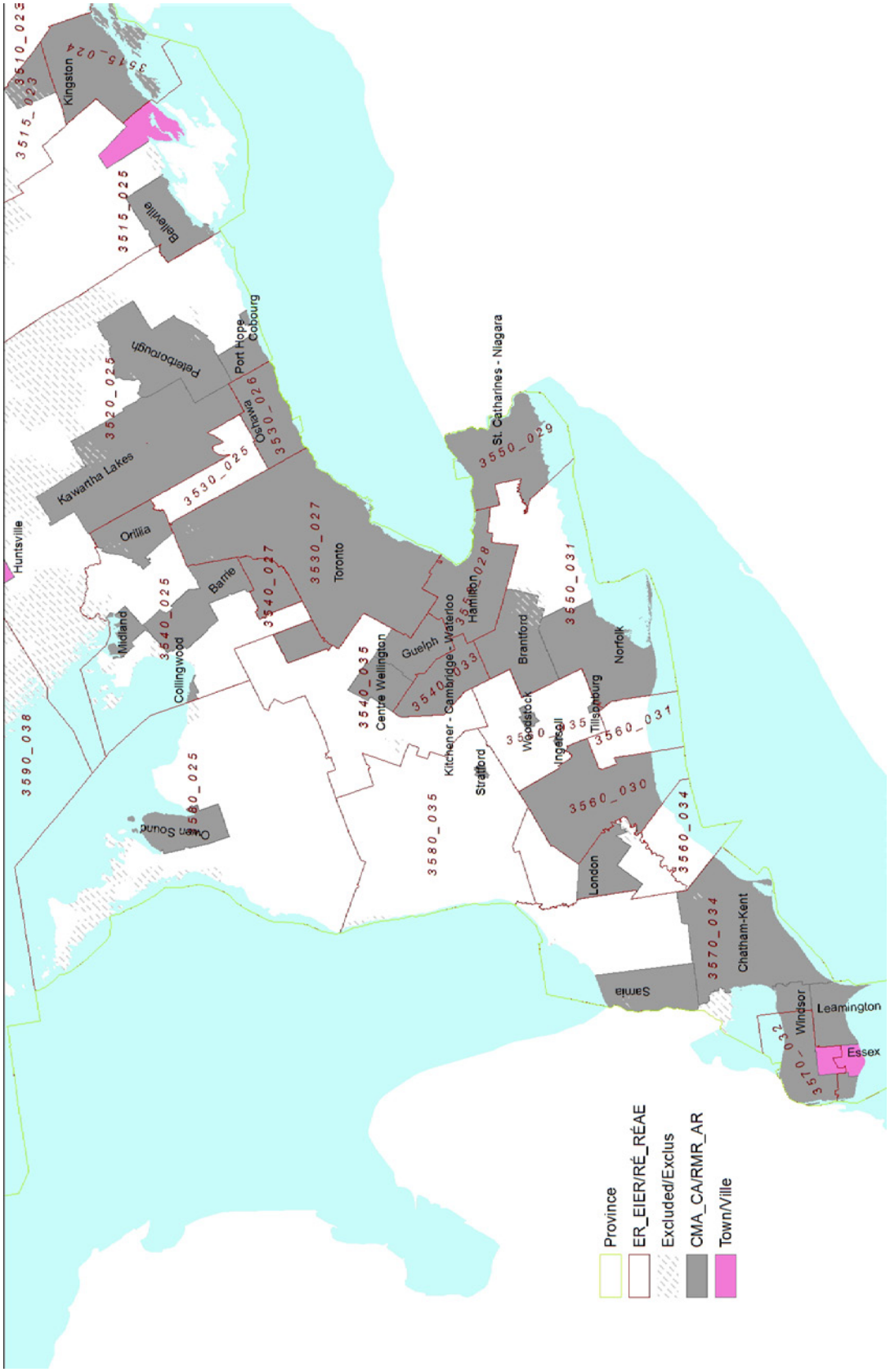
**Map 6a – Ontario East**
**CMACA, EI-EIER Intersection Boundaries**



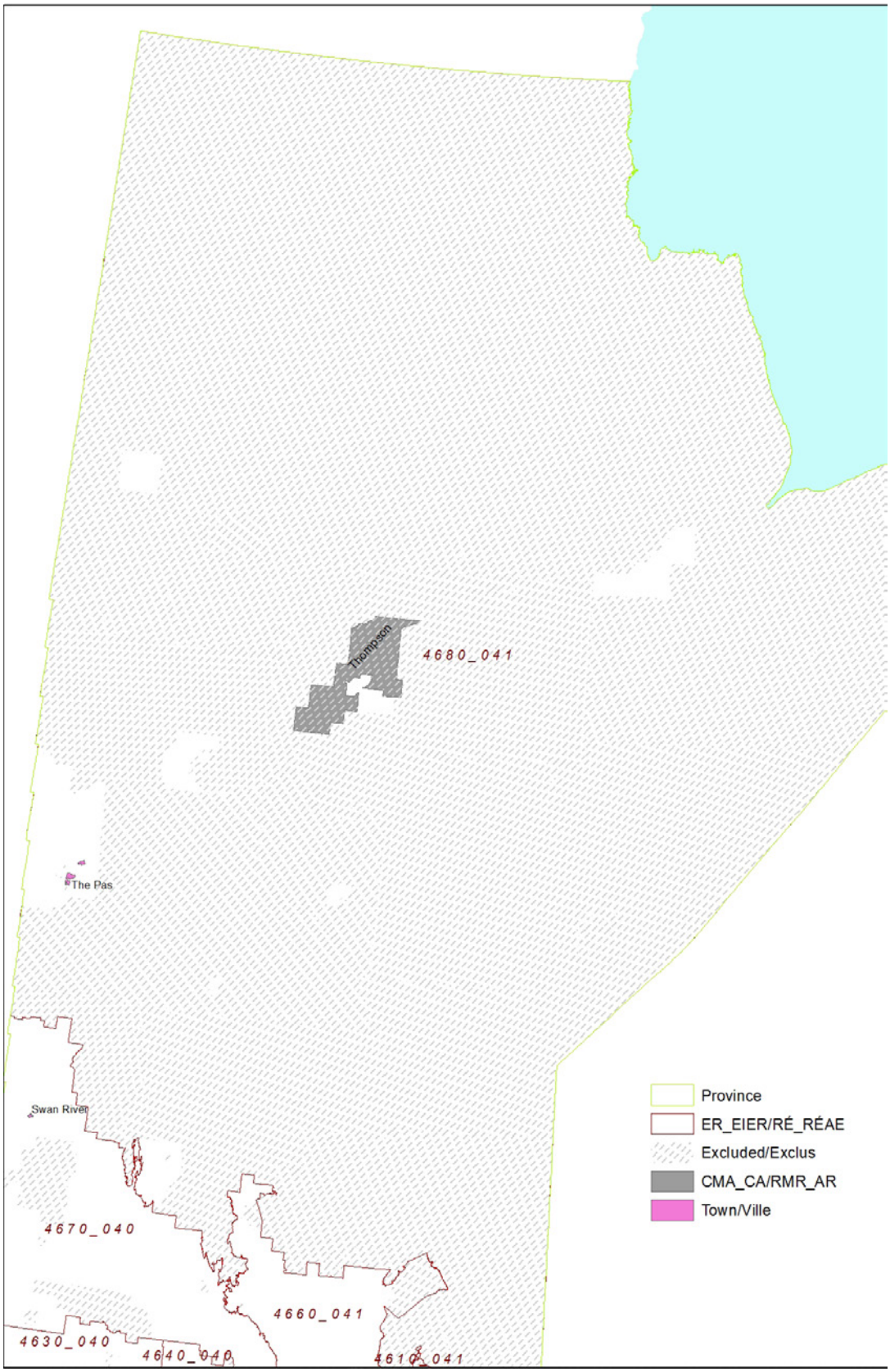
**Map 6b – Ontario North**
**CMACA, EI-EIER Intersection Boundaries**



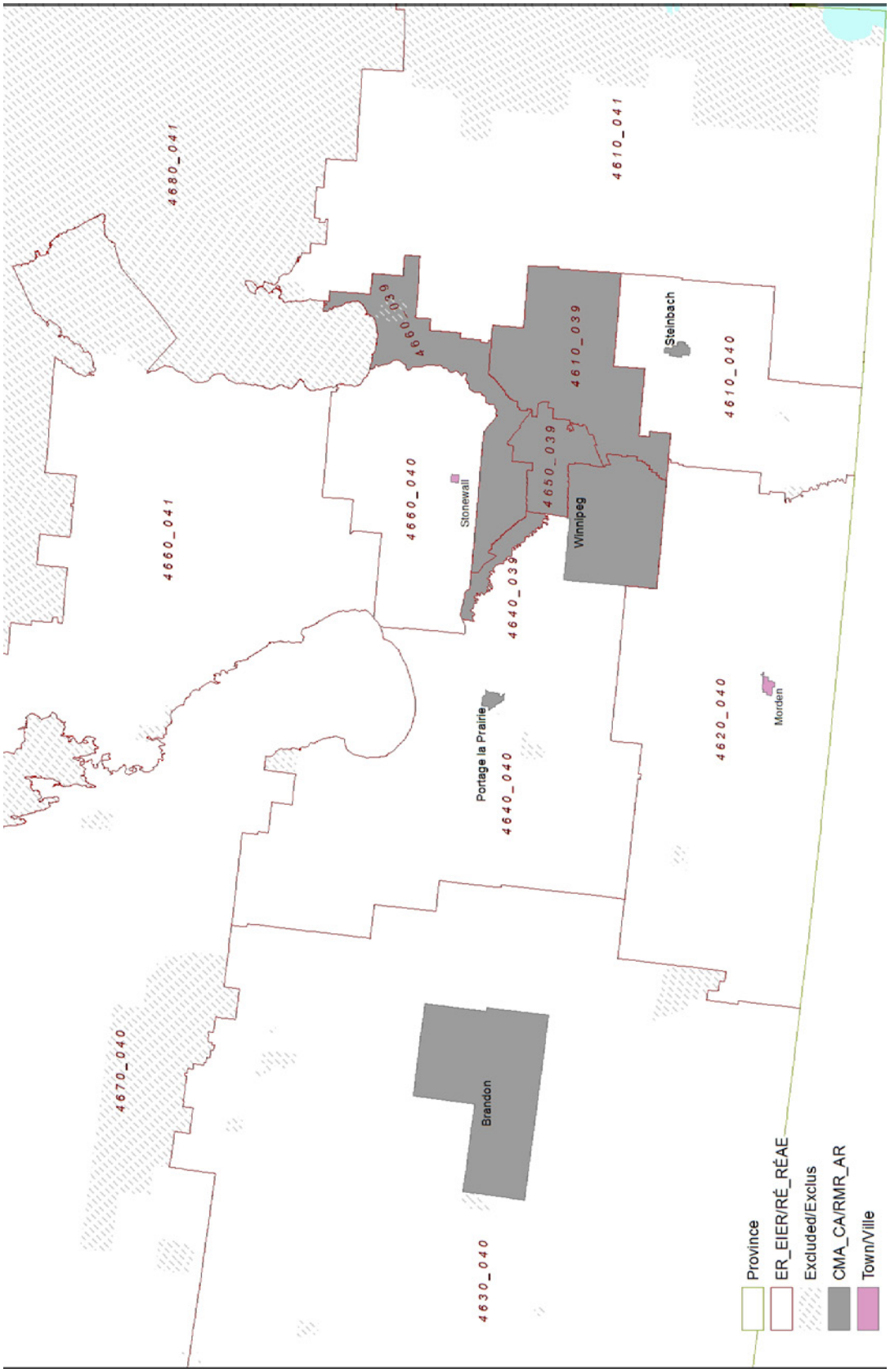
**Map 6c – Ontario South**
**CMACA, EI-EIER Intersection Boundaries**



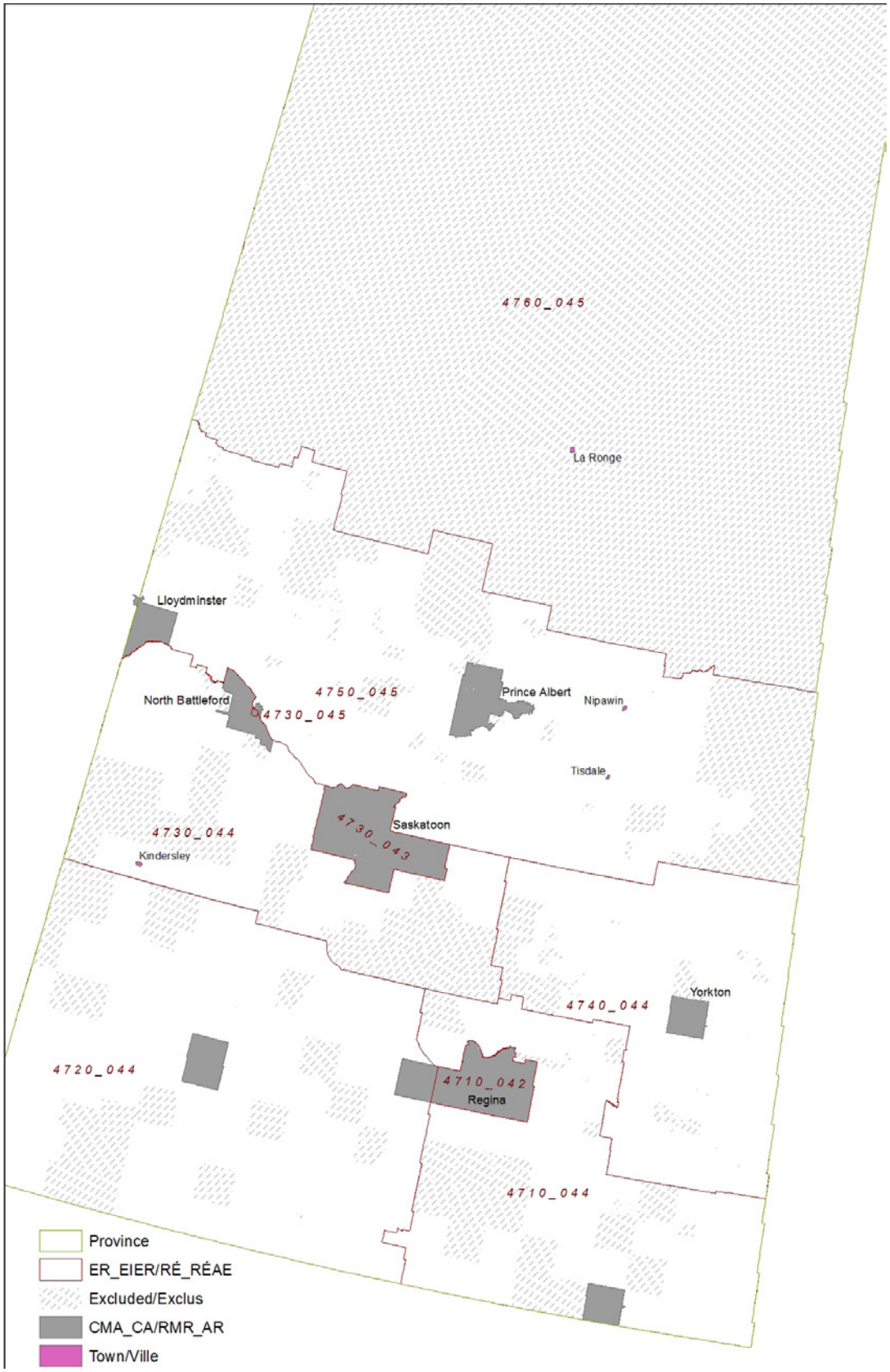
**Map 7a – Manitoba North**
**CMACA, EI-EIER Intersection Boundaries**



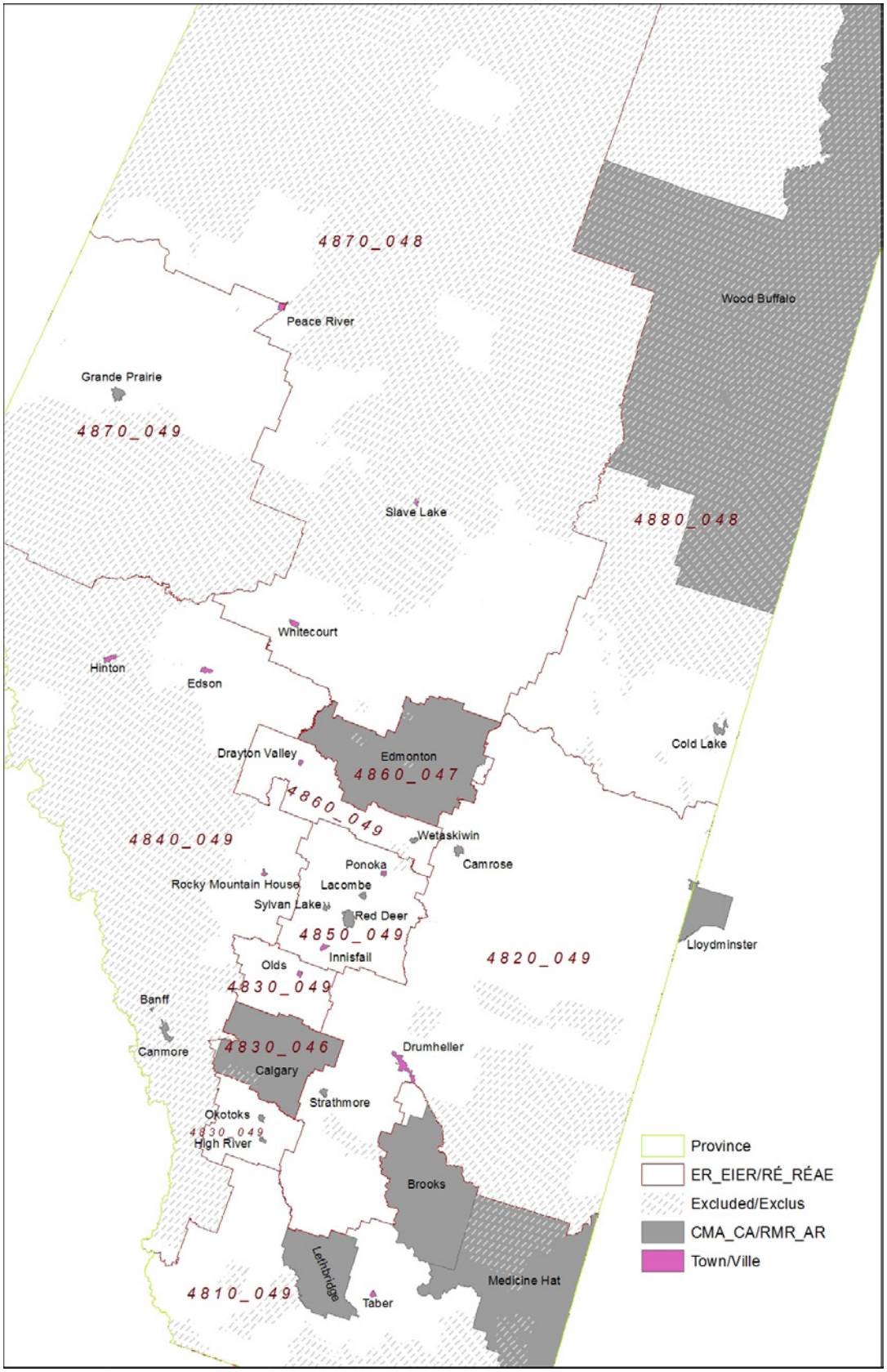
**Map 7b – Manitoba South**
**CMACA, EI-EIER Intersection Boundaries**



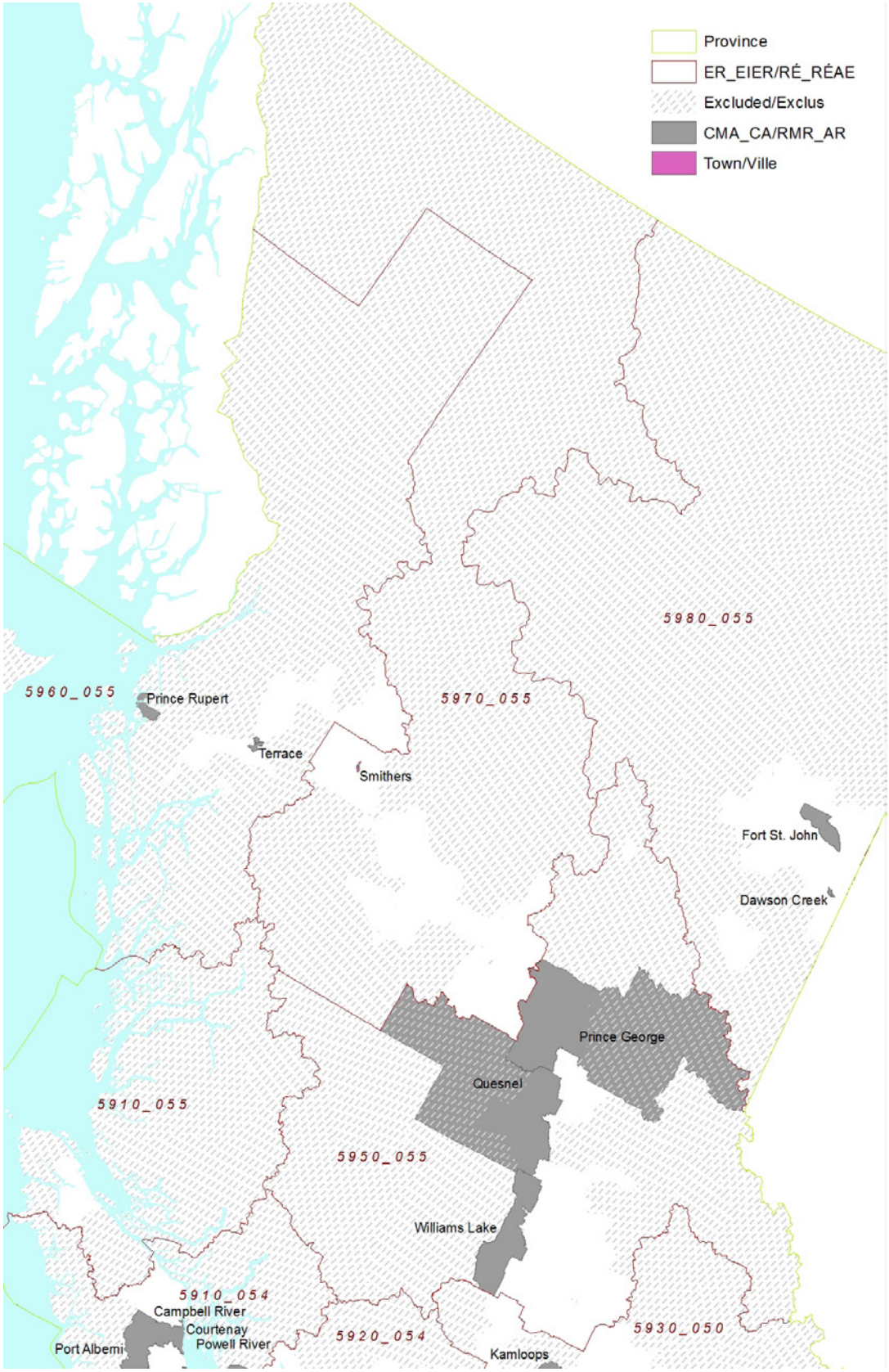
**Map 8 – Saskatchewan**
**CMACA, EI-EIER Intersection Boundaries**



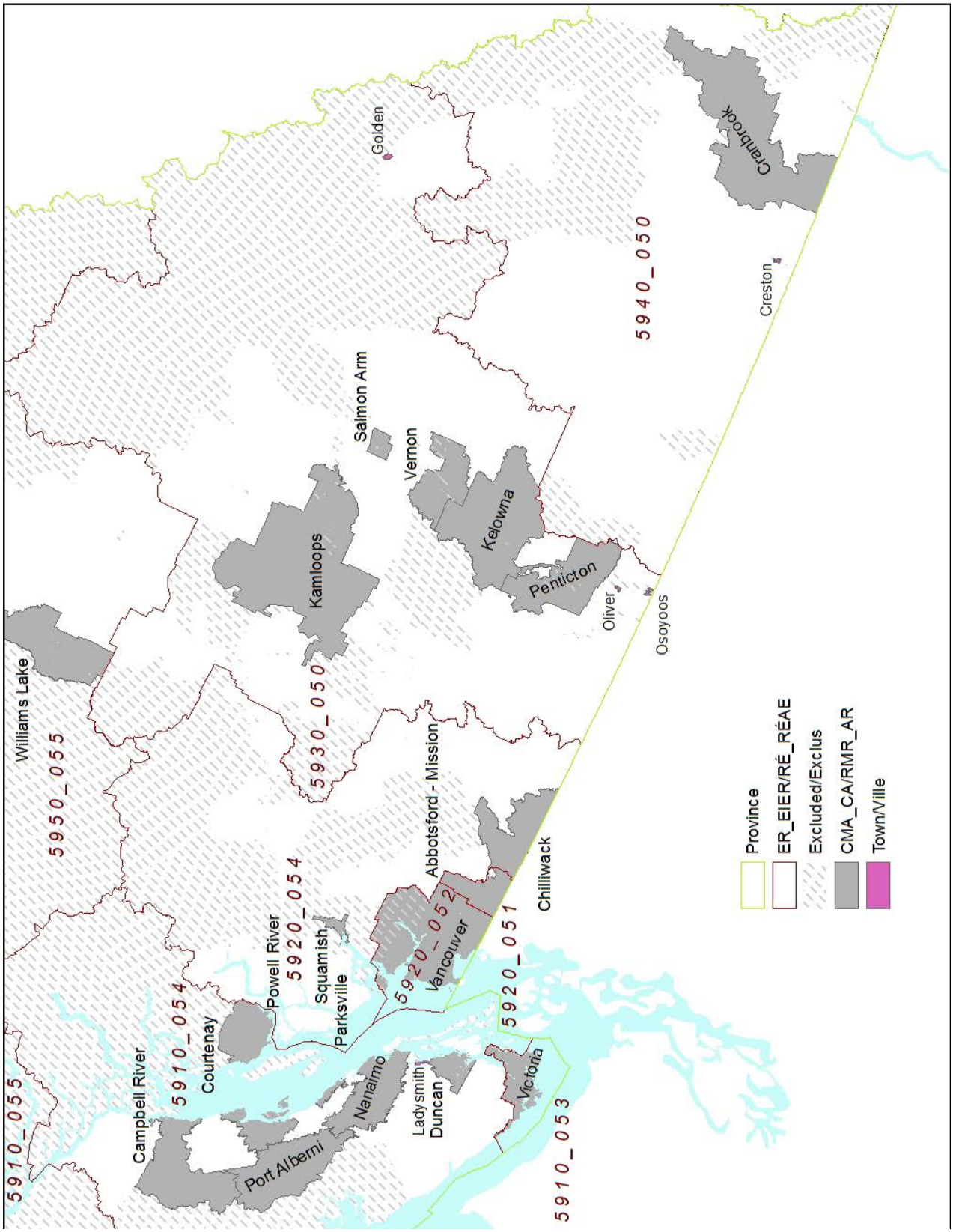
**Map 9 – Alberta**
**CMACA, EI-EIER Intersection Boundaries**



**Map 10a – British Columbia North**
**CMACA, EI-EIER Intersection Boundaries**



**Map 10b – British Columbia South**
**CMACA, EI-EIER Intersection Boundaries**



# Appendix F Definition of variables used to form imputation classes

## Age groupings

| AGEGP1 | AGEGP2 | AGEGP3 | Age Range |
|---|---|---|---|
| 1 | 1 | 1 | 15 to 19 |
| 2 | 2 | 2 | 20 to 24 |
| 3 | 3 | 2 | 25 to 29 |
| 3 | 4 | 2 | 30 to 34 |
| 3 | 5 | 2 | 35 to 44 |
| 3 | 6 | 2 | 45 to 54 |
| 4 | 7 | 2 | 55 to 64 |
| 5 | 8 | 3 | 65+ |

## Occupation groupings

| OCC4 | OCC10 | Description |
|---|---|---|
| 01 | 01 | Management, business, finance, and natural and applied sciences |
| 02 | 02 | Administrative and clerical |
| 02 | 03 | Health |
| 02 | 04 | Education, law, and social, community and government services |
| 03 | 05 | Art, culture, recreation, and sport |
| 03 | 06 | Sales and services |
| 03 | 07 | Trades, transport, and equipment operators |
| 03 | 08 | Natural resources and agriculture |
| 03 | 09 | Manufacturing and utilities |
| 04 | 10 | Never worked before or last worked more than 1 year ago or permanently unable to work |

## Education grouping

| EDUC | Description |
|---|---|
| 0 | Person does not have a high school diploma |
| 1 | Person does have a high school diploma |

## Class of worker

| COW | Description |
|---|---|
| 1 | Paid employee |
| 2 | Self employed |
| 3 | Unpaid family worker |

## Student status

| STUD | Description |
|---|---|
| 0 | Not a full-time student |
| 1 | Full-time student |

## Dwelling owned or rented

| DWELRENT | Description |
|---|---|
| 1 | Dwelling owned |
| 2 | Dwelling rented |

### Province of residence

| PROV | Province |
|---|---|
| 10 | Newfoundland and Labrador |
| 11 | Prince Edward Island |
| 12 | Nova Scotia |
| 13 | New Brunswick |
| 24 | Quebec |
| 35 | Ontario |
| 46 | Manitoba |
| 47 | Saskatchewan |
| 48 | Alberta |
| 59 | British Columbia |

### Sex

| SEX | Description |
|---|---|
| M | Male |
| F | Female |

### Labour force status path options (Temporary path)

| TPATH | LFSSTAT | Description |
|---|---|---|
| 1 | 1 | Employed and at work |
| 2 | 2 | Employed and away from work |
| 3 | 3 | Temporarily laid off |
| 4 | 4 | Unemployed, job seeker |
| 5 | 5 | Unemployed, future start |
| 6 | 6 | Not in the labour force |
| 7 | 7 | Permanently unable to work |
| 8 | 2, 3, 4, 5 or 6 | |
| 9 | 3, 4, 5 or 6 | |
| 10 | 2, 4, 5 or 6 | |
| 11 | 4, 5 or 6 | |
| 12 | 5 or 6 | |

### Did the respondent have more than one job or business last week?

| MULTJOB | Response |
|---|---|
| 1 | Yes |
| 2 | No or no response |

### Country of birth

| IMM | Description |
|---|---|
| 1 | Canada |
| 2 | United States |
| 3 | Other |

### Aboriginal identity

| ABQ1 | North American Indian, Métis, or Inuit |
|---|---|
| 1 | Yes |
| 2 | No |

### Last month's labour force status groupings

| LMLFS3 | LMLFS7 | Description |
|---|---|---|
| 1 | 1 | Employed and at work |
| 1 | 2 | Employed and away from work |
| 2 | 3 | Unemployed, temporarily laid off |
| 2 | 4 | Unemployed, job seeker |
| 2 | 5 | Unemployed, future start |
| 3 | 6 | Not in the labour force |
| 3 | 7 | Permanently unable to work |

### Last month's industry group

| LMINDG | Description |
|---|---|
| 1 | Agriculture, Forestry, Fishing and Hunting |
| 2 | Mining and Oil and Gas Extraction |
| 3 | Utilities |
| 4 | Construction |
| 5 | Manufacturing |
| 6 | Wholesale Trade |
| 7 | Retail Trade |
| 8 | Transportation and Warehousing |
| 9 | Information and Cultural Industries |
| 10 | Finance and Insurance |
| 11 | Real Estate and Rental and Leasing |
| 12 | Professional, Scientific and Technical Services |
| 13 | Management of Companies and Enterprises |
| 14 | Administrative and Support, Waste Management and Remediation Services |
| 15 | Educational Services |
| 16 | Health Care and Social Assistance |
| 17 | Arts, Entertainment and Recreation |
| 18 | Accommodation and Food Services |
| 19 | Other Services (except Public Administration) |
| 20 | Public Administration |

# Appendix G Composite auxiliary variables

The following is a list of the 28 composite auxiliary variables that are used for LFS composite calibration. These variables are defined at the province level. An asterisk (*) indicates that the auxiliary variable does not need to be specified because it can be deduced from other auxiliary variables.

### Labour force characteristics of previous month (no breakdown)

Employed, 15+
Unemployed, 15+
** Not in the labour force, 15+*

### Labour force characteristics of previous month by age/sex groups

Employed males, 15 to 24
Unemployed males, 15 to 24
Not in labour force males, 15 to 24

Employed males, 25+
Unemployed males, 25+
Not in labour force males, 25+

Employed females, 15 to 24
Unemployed females, 15 to 24
Not in labour force females, 15 to 24

** Employed females, 25+*
** Unemployed females, 25+*
** Not in labour force females, 25+*

### Employment of previous month by industry

Employed in natural resources, 15+
Employed in utilities, 15+
Employed in construction, 15+
Employed in manufacturing, 15+
Employed in trade, 15+
Employed in transportation and warehousing, 15+
Employed in finance, insurance and real estate, 15+
Employed in professional, scientific and technical services, 15+
Employed in management, administrative and other support, 15+
Employed in educational services, 15+
Employed in health care and social assistance, 15+
Employed in information, culture and recreation, 15+
Employed in accommodation and food services, 15+
Employed in other services, 15+
Employed in public administration, 15+
** Employed in agriculture, 15+*

### Employment of previous month by class of worker

Employed, public sector employee, 15+
Employed, privates sector employee, 15+
** Employed, private sector, self-employed, 15+*