Catalogue no. 71-526-X

Methodology of the Canadian Labour Force Survey

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Client Services, Labour Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-866-873-8788, email: labour@statcan.ca).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.ca or contact us by e-mail at infostats@statcan.ca or by telephone from 8:30 a.m. to 4:30 p.m. Monday to Friday:

Statistics Canada National Contact Centre

Toll-free telephone (Canada and the United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository services program

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

Information to access the product

This product, Catalogue no. 71-526-X, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. 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.ca under “About us” > “Providing services to Canadians.”

Statistics Canada

Methodology of the Canadian Labour Force Survey

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2008

All rights reserved. The content of this electronic publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it be done solely for the purposes of private study, research, criticism, review or newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form, by any means—electronic, mechanical or photocopy—or for any purposes without prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2008

Catalogue no. 71-526-X
ISBN 978-0-662-47995-6

Frequency: occasional

Ottawa

La version française de cette publication est disponible sur demande (nº 71-526-X au catalogue).

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 cooperation and goodwill.

Symbols

The following standard symbols are used in Statistics Canada publications:

. not available for any reference period

.. not available for a specific reference period

… not applicable

0 true zero or a value rounded to zero

0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded

^{p} preliminary

^{r} revised

x suppressed to meet the confidentiality requirements of the *Statistics Act*

^{E} use with caution

F too unreliable to be published

Acknowledgement

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

Of particular note are those who were the primary authors of various chapters: Jean-François Beaumont, René Boyer, Guy Laflamme, Danielle Lebrasseur, John Lindeyer and Claude Turmelle. These people worked on different aspects of the 2004 redesign of the methodology of the Labour Force Survey, and thus were the best people to write about it.

Besides these, many other 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

Chapter 1 Introduction and overview

Introduction

This publication is a reference work on the methodology of the Labour Force Survey (LFS), which covers all of Canada. This work will focus on the methodology used for the ten provinces. It describes the changes made during a recent sample redesign and covers all survey steps.

A separate document called the *Guide to the Labour Force Survey* (available on line at www.statcan.ca) is a complement to this report and highlights 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 Inter-departmental 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 rose, since users wanted a broader range of labour market statistics, in particular more detailed regional data. The range of estimates produced by this survey has grown considerably over the years, and today 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 (paid work, self employment, full- and part-time work) and unemployment. The main monthly indicators published include the unemployment rate, the employment rate and the participation rate. It is one of the main sources of information on the individual characteristics of the working-age population (such as age, marital status, level of education and family status).

Employment estimates are broken down at various levels, such as industry, occupation, job tenure, and the usual and actual hours worked. Some of the questions asked make it possible to examine a wide variety of topical issues, such as involuntary part-time employment, multiple job-holding, and absence from work.

Unemployment estimates are produced by industry and occupation, duration of unemployment, type of work sought, and activity before looking for work. 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 subprovincial regions, such as Employment Insurance Economic Regions (EIERs) and Census Metropolitan Areas (CMAs). In the past few years, standard labour market indicators, such as census divisions (CD) and Canada Employment Centres, have been computed using special estimation techniques. 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 on line). A variety of labour market data are also available through CANSIM, the Statistics Canada database and electronic extraction system. With more than 1,000 chronological series, this database is updated monthly with LFS data.

The LFS is also the source of the *Labour Force Historical Review* CD-ROM (Catalogue No. 71F0004X), which contains detailed data in cross-sectional and chronological series from 1976 to the current year. The survey can produce much more information than what is periodically published. Special tabulations are produced on a cost-recovery basis. For more information about the survey products and services, please see Section 8 of the *Guide to the Labour Force Survey*.

1.3 General survey overview and document structure

In the provinces, the LFS covers 98% of the population. People living on First Nation reserves and Crown lands, residents of institutions and certain remote regions, as well as full-time members of the Canadian Forces are excluded from the survey target population because collecting their data would bring about operational problems. 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 are also excluded because most are unable to participate in the labour market.

The survey uses a two-stage sample design[1]. In the first stage, we select a sample of geographical regions, which are called primary sampling units (PSUs). In each PSU selected, we draw a sample of dwellings in the second stage. The dwellings selected remain in the sample for a period of six months. Each month, the dwellings that participated in the survey for six months, or 1/6 of the sampled dwellings, are replaced by other dwellings. As a result, there is an overlapping of 5/6 of the respondents in two consecutive months. This approach helps to improve the quality of the change estimates between two months and to control the response burden imposed on each dwelling. Chapters 2 and 3 provide more information on the sample design.

Data collection for the LFS takes place in the week following the reference week. Usually, the reference week is around the 15th day of the month. In 2004, around 51,000 households responded to the LFS questionnaire every month. The LFS interview only takes an average of eight minutes and the data are collected using a computer-assisted interviewing system. To reduce travel, and thus reduce collection costs, we try to do most interviews by telephone. More information on the collection strategy is presented in Chapter 4.

In the days following collection, we process, impute and weight the data and derive the quality indicators. These steps are described in Chapters 5, 6 and 8. Despite the volume of information to process every month, Statistics Canada publishes the LFS estimates only 13 days after the end of the interview week.

It is important to note that the LFS sampling frame is also used for most of Statistics Canada's social surveys. This is covered in Chapter 9. Several appendices, each dealing with a special topic, and other survey materials, are included at the end.

1.4 Changes introduced in November 2004

Every 10 years after the Census of Population, we redesign the LFS in order to account for the evolution of the population's characteristics and the needs of data users and to update the geographical information required to carry out the survey. Unlike the redesign in 1994, the questionnaire and collection application were not modified this time around.

The 2004 redesign was developed in the context of a restricted budget framework. As a result, some studies carried out during previous redesigns could not be updated. We had to presume that the conclusions of the previous redesigns remain valid. This budget constraint also meant that the redesign had to be implemented without having to change existing computer systems. Lastly, part of the redesign had to be funded by reducing the survey sample size for a period of three years. This reduction strategy is described in Section 2.5.3.

The new sample design was gradually introduced beginning in November 2004. To reduce survey operating costs, two major changes to the methodology were introduced.

Before November 2004, the first of six interviews was conducted in person. To reduce collection costs, the first interview is now done by telephone for certain dwellings in urban centres. This strategy is described in Chapter 4.

The second change is aimed at reducing the transition cost from the previous sample design to the new one. A significant part of a redesign budget is earmarked for a list of addresses for each PSU selected in the new design. To reduce the costs of this listing operation[2] and improve the survey frame coverage, we used the Address Register (AR), a database containing the addresses of dwellings in urban centres. More information on this change is given in Chapter 3.

In addition to these two major changes, other improvements were made to the sample design. In the past, the Statistics Canada geographical database did not cover the entire territory of all 10 provinces. This database is required to establish the boundaries of the PSUs. Consequently, before the 2004 redesign, some of the PSUs were defined to satisfy specific survey needs. For the rest of the territory, we had to use the 1991 enumeration areas[3] as PSUs. The quality of the geographical database has greatly improved since the 1994 redesign. For the first time in 2004, we were able to define the boundaries of PSUs for the entire territory of the ten provinces. The PSU creation strategy is presented in Section 2.6.4. We also developed an application to draw maps for data collection in each selected PSU. The geographical database is the cornerstone of this application.

1. A three-stage design is used only to survey average-sized isolated urban centres.

2. A listing operation consists of compiling a list of residential addresses in the PSU. The interviewer uses a map identifying the PSU boundaries.

3. Small area composed of one or more neighbouring blocks, used by Statistics Canada for distributing questionnaires to households and dwellings (census collection). All of Canada is divided into enumeration areas.

To better control the sample distribution, and in turn, collection costs, we implemented a specific strategy for regions with a high collection cost (see Section 2.5.2 and Chen, Lindeyer and Laflamme 2004). We also introduced methods to target the immigrant population in large centres and the aboriginal population in the four Western provinces (see Section 2.6.4.2). To decrease the maintenance costs associated with the survey frame, the new sample design no longer contains a survey frame of apartments. Lastly, the sample of small rural areas is now selected using a two-stage design rather than the three-stage design used in the past.

Chapter 2 Sample design

The sample design describes all the steps to be carried out when selecting a sample of persons. It aims to improve the quality of the estimates produced and to control costs. Various strategies are in place to achieve this objective.

A significant part of a survey's budget is earmarked for data collection. In addition, the sample design has little impact on the other budgetary items of a survey (*e.g.*, processing and dissemination). Therefore, in practice, the sample design tries to reduce collection costs while maximizing data quality.

Section 2.1 describes the LFS sample design. Section 2.2 presents some basic concepts of survey theory that will be used throughout this chapter. Readers familiar with these concepts can move on to the following sections. Section 2.3 provides the AR usage strategy, and the rest of the chapter is devoted to four techniques used to improve the effectiveness of the sample design.

2.1 Description of the sample design

It is impossible to contact everyone 15 years of age and over every month to determine their employment status in order to produce the necessary estimates. In addition, there is no administrative source to produce these estimates, which means we have to produce them using a sample of persons.

We cannot directly select a sample of persons to interview because we do not have a complete list of the persons residing in the ten provinces. However, even if such a list did exist, the persons selected would be spread out. As a result, travel costs would be exorbitant if the interviewer had to go to certain respondents in person.

To reduce travel costs, the sample of persons is taken through two consecutive selection stages. This selection method is called two-stage sampling[4]. In the first stage, we select a sample of geographical regions, called primary sampling units (PSUs). For each PSU selected, we draw up a list of dwellings. Then we select a second-stage dwelling sample from these lists. All the residents of the dwellings selected in the second stage make up the LFS sample of persons. This selection method reduces geographic spread of the sampled persons and prevents the necessity of creating a list of all the addresses in the ten provinces.

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. Furthermore, when the overlap is increased, the burden imposed on respondents also rises because they must participate in the survey several times. In turn, this increase in burden could lead to an increase in the nonresponse rate. Therefore, in the end, increasing the overlap has an adverse effect on the response burden. It then becomes necessary to establish a compromise between the quality of the change estimates and the burden imposed on respondents. We mention that the collection costs are also influenced by the extent of the overlap. For example, it costs more to obtain a response in the first month than in the subsequent months. Therefore, a bigger overlap reduces survey operation costs.

By considering these factors, a decision was made that each dwelling will remain in the LFS sample for six consecutive months. Subject to this limitation, the maximum overlap of the sample between two consecutive months is 5/6; therefore, it is necessary to replace 1/6 of the sample of dwellings each month. To implement this strategy, the LFS dwelling sample is divided into six rotation panels or groups. Each group represents the population observed. An initial contact with the dwellings in the first rotation group was made in January. These dwellings will remain in the sample until June inclusively. In July, all the dwellings in Group 1 will be replaced by new dwellings. The second group is made up of the dwellings surveyed from February to July inclusively, and so on for the other rotation groups. More information about the rotation of the dwelling sample is provided in Section 2.7.2.

Managing the LFS sample through rotation groups is a simple method for selecting samples for other Statistics Canada surveys. Since each rotation group represents the population, we can 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 to select samples for other household surveys is given in Chapter 9.

We mention that overlapping the LFS monthly surveys opens the door to more effective processing and estimation methods. The methods currently used by the LFS are described in Chapters 5 and 6.

4. A three-stage design is used on occasion to process average-sized isolated urban centres. This method is presented in Section 2.6.4.

2.2 Some basic survey theory concepts

This section presents some concepts required to understand the description of the sample design provided in this chapter. A conceptual overview of survey theory is available in Satin and Shastry (1992). For more information on this theory, readers are invited to consult one of the many books dealing with this subject (*e.g.*, Cochran 1977 or Särndal, Swensson and Wretman 1992).

The LFS estimates are produced from a probabilistic sample, *i.e.*, a sample for which the persons to survey are selected at random. It is important to note that the estimate produced from a sample differs from the resulting estimate if the entire population was interviewed. This difference is sampling error. Similarly, the estimate produced from a given sample differs from the estimate that would be produced using another sample. In addition to sampling errors, the results of a survey also have non-sampling errors. Chapter 8 presents the procedures in place to control this type of error.

Two important error measurements are bias and sampling variance. These two concepts can be defined intuitively. Let us suppose that it is possible to select several samples based on a given sample design. For each sample, we produce an estimate of the characteristic of interest (*e.g.*, the number of unemployed, average personal income, *etc.*). The estimate is biased if the average of the estimates produced from all the possible samples differs from the estimate we would obtain by surveying all the individuals of the population. In addition, the variance between these different estimates is the result of sampling.

Biased estimates can be caused by a number of sources, such as an imperfect survey frame, the method used to produce the estimate, and persons' nonresponse. This error component can be difficult to measure in practice. It also has little effect on the definition of the sample design. Consequently, we ignored the potential bias when implementing the first four steps of the sample design documented in this chapter.

Sampling variance measures the spread between the estimates produced from all the possible samples: the smaller the sampling variance, the more accurate the estimate. An estimate of sampling variance can be obtained from one probabilistic sample.

Other measures are derived from sampling variance, including standard error. This measure 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. Another measure is the coefficient of variation (CV), which is a relative measure of the quality of an estimate. By definition, the CV is the standard error divided by the estimate. Once again, the smaller the CV, the more accurate the estimate is. A third measure is design effect, which is a relative measure used to compare the effectiveness of one sample design to another. More information on these measures is given in Chapter 8.

Sampling variance is a measure of the design's efficiency. When designing the sample, we try to reduce the sampling variance. From another viewpoint, a more efficient sample design, or one that results in a smaller sampling variance, helps to reduce the sample size compared to another less efficient design, while maintaining the quality of the estimates.

Several factors influence the sampling variance of an estimate. The most important are the number of persons in the population, the number of persons in the sample, the sample design used to draw the sample, the response rate, and the homogeneity of the characteristic of interest in the population. The sample allocation (see Section 2.5) ensures that the number of dwellings selected is sufficient to produce various survey estimates of adequate quality. Stratification (see Section 2.6) helps to improve the sample design's effectiveness by grouping together similar dwellings. In a two-stage survey such as the LFS, homogeneity of the characteristic of interest within the PSU and in the dwelling also influences the sampling variance. The more homogeneous the dwellings in the same PSU, the less effective the sample design. The same logic applies to homogeneity within dwellings.

Meanwhile, the creation of PSUs helps to reduce collection costs. This is discussed in Section 2.4. Finally, certain sample selection methods help to improve the effectiveness of the sample design. Section 2.7 describes the selection method used by the LFS.

By controlling each of these parameters, it is possible to define an effective sample design, *i.e.*, one with a small sampling variance for a given sample size and a set operating cost.

2.3 The Address Register and its impact on the sample plan

The Address Register (AR) is a database containing the address of dwellings in urban centres (see Chapter 3.1 for a description of the AR). During the redesign, we considered the possibility of selecting a sample of

dwellings directly from the AR. In an urban area, this method could have been more economical and efficient than the two-stage sample design traditionally used by the LFS. However, there are three major problems associated with this.

First, the quality of the AR is not geographically uniform. It would have been difficult to define the regions from which we could have selected a sample of dwellings directly from the AR. In addition, when the sample was being redesigned, the AR updating process was uncertain[5]. In this context, it would have been difficult to maintain the quality of the survey frame over time. Finally, selecting dwellings directly from the AR would have required significant changes to the existing computer systems. However, the redesign budgetary framework did not make it possible to do this. For these reasons, it was decided that selecting dwellings directly from the AR was not advantageous for this redesign.

However, it was still preferable to use the AR to reduce collection costs and to improve survey coverage. The compromise consisted of using the AR as a tool to improve the effectiveness of the address listing operation in the selected PSUs. For each PSU selected, we determined the quality of the information available in the AR. Based on this measurement, we selected the appropriate method for using the register. This strategy is described in Chapter 3.

Using the AR helps to reduce listing costs, but does not reduce the other collection cost components (travel to conduct an interview, time required to conduct an interview, *etc*.). We mention that using the AR does not require us to change any procedures traditionally used to update the survey frame. In this context, we take advantage of the benefits of using the AR without significantly disrupting the work in the field.

Based on this strategy, using the AR has little effect on the development of the other sample design components, but directly affects the collection cost structure, and in turn, the PSU creation strategy. However, this impact was not known during the redesign.

2.4 Creation of PSUs

The first step of the sample design consists of defining the PSUs to be used for the first-stage sample selection. Once defined, they form the survey area frame. In Section 2.6, we will see that it is possible to structure the survey frame to increase the design's efficiency.

There are standard geographical units defined in the Standard Geographical Classification. To reduce redesign costs, we assessed the possibility of using one of these units to define the PSUs. Unfortunately, no standard unit satisfied the survey needs; some were too small, while others were too large. Furthermore, it would have been difficult to correct these problems. This is why we had to create the PSUs for the entire territory of the ten provinces. We were able to do this because the geographical database now covers the entire territory. This alternative was inconceivable during the previous redesign.

The size and shape of the PSUs directly affect the sample design's efficiency as well as the collection costs. With regard to design efficiency, it is preferable to select few dwellings from small PSUs. If we push this argument to the extreme, the ideal solution would be to create PSUs that contain only one dwelling and to select this dwelling when the corresponding PSU is selected. This situation is like selecting dwellings directly from a dwelling list. Unfortunately, this solution is not possible with an area frame.

The size of the PSUs and the number of households selected per PSU affect several facets at the same time. These effects often influence several components of the design. For example, when we reduce the number of dwellings selected per PSU, we must select more PSUs to attain the desired sample size. As a result, the selected dwellings are generally further away from one another, which increases the travel costs for in-person interviews. Similarly, we must list more PSUs. Furthermore, the size of the PSU has an impact on the listing costs; the bigger the PSU, the harder and more costly it is for the interviewer to draw up a list of dwellings. By contrast, a large PSU can remain in the sample for longer before being replaced. The listing cost of a large PSU is therefore amortized over a longer period.

To determine the ideal size of a PSU and the optimal number of dwellings to select per PSU, two elements are necessary. First, we must have a tool to evaluate the sampling variance resulting from different scenarios. This tool can be built using census data. We must also have a relatively accurate model of collection costs that helps to estimate the costs for different scenarios regarding PSU size and the number of dwellings selected per PSU. To build this model, we must have detailed information on costs. With the introduction of the first contact by telephone (see Section 4.3) and use

5. The AR is mainly used to help collect data for the Census of Population. A process adapted to the census needs is in place to update the AR.

of the AR, it was virtually impossible to build a valid cost model. In this context, we were unable to re-evaluate the ideal size of the PSUs and the optimal number of dwellings to select from the sampled PSUs. Consequently, the conclusions of the last redesign were retained.

The size of the PSUs was set at 200 occupied dwellings. Other criteria were established to frame their creation. To reduce collection costs, it is preferable to create compact PSUs. A PSU is said to be compact if the perimeter-surface ratio is small. In practice, we want to avoid elongated PSUs in order to reduce the distance to travel during collection. To also reduce the number of kilometres to travel, it is advantageous for all the roads of the PSU to be accessible without having to leave the unit boundaries (*e.g.*, PSUs should not overlap a highway or river). It is also desirable for PSUs to respect the regional boundaries for which we want to produce estimates.

The PSUs were created using the Geographic Area Delineation System (GArDS), a system originally developed by the Geography Division of Statistics Canada to meet census needs. Since the needs of the LFS are similar to those of the census, the expertise gained with the census could be used for the LFS.

The LFS PSUs are created by grouping together census blocks[6]. To do this, GArDS began with an initial solution and tried to improve it by moving the census blocks from one PSU to another. Each solution was compared to the previous one using a measurement combining various criteria mentioned earlier. If this measurement was less than that of the previous solution, it was kept. GArDS explores permutations until it is impossible to improve the solution found or the time allotted for research is up.

The PSUs created by GArDS cannot be directly used by the LFS. In some circumstances, these PSUs may sometimes have too many or too few dwellings and we must combine small PSUs. These small PSUs are caused by operational limitations of GArDS. At the other extreme, some census blocks have more than 200 dwellings. PSUs that include these large census blocks will, by definition, be too big. These large PSUs were divided using information from the AR. Finally, a manual validation step was carried out for a PSU sample and some corrections were made to the PSU boundaries.

We also had to do special processing for PSUs in remote regions. These regions are characterized by large, uninhabited areas. In the past, a particular unit in the Standard Geographical Classification was used to create the PSUs in these regions. However, this unit no longer exists in the 2001 Classification. Therefore, it was necessary to implement a specific strategy for these regions.

When left alone, GArDS ensured that each parcel of territory was assigned to a PSU. In the remote regions, this approach led to huge PSUs that covered several hundred square kilometres. Using these would have generated enormous collection costs and frustrated interviewers, who would have had to travel several hundred kilometres to contact respondents. To ensure that the PSUs covered small areas, those created by GArDS were modified to exclude the parcels of territory with a small population. The PSUs were then redefined to only cover parcels of land with a relatively high population.

Once the PSUs were created, a detailed analysis was done to identify those that would have a very high collection cost. This analysis helped to slightly adjust the allocation and stratification strategies in order to control the collection costs in these PSUs. These strategies are described in Sections 2.5.2 and 2.6, respectively.

2.5 Sample allocation

As explained in Section 2.2, the number of dwellings sampled has a direct impact on the quality of the estimates that can be produced by the survey. Since the LFS produces estimates at various geographical levels (Canada, provinces, economic regions, *etc.*) and for several variables, it is necessary to reach a valid compromise for all these estimates when allocating the sample.

This search for a compromise is not simple since it depends on diametrically opposed needs. For example, the optimal allocation to produce estimates at the national level would be approximately proportional to the population of each province. However, this kind of allocation would produce poor estimates for the provinces with small populations, and excellent estimates for provinces with the most residents. Therefore, this approach was not selected. Furthermore, an allocation that ensures a standard quality for each province would produce poorer estimates at the national level. Once again, this was not an appropriate solution.

6. An area equivalent to a city block bounded by intersecting streets. These areas cover all of Canada.

The sample allocation specifies the number of dwellings to select in each geographical level. It is established to ensure that the sample can produce estimates that satisfy the pre-specified quality objectives. This is a crucial step because the subsequent steps depend on it and it ensures that the survey resources are effectively used. Too many dwellings assigned to a given region will produce estimates of better quality than that established in the survey objectives. However, this surplus of sampled dwellings in a region is obtained to the detriment of the other regions since the survey budgetary envelope is fixed. The sample allocation is made to meet all of the needs as well as possible while maintaining overall efficacy of the survey design.

We mention that the LFS sample allocation aims to meet the objectives of this survey. However, the sample is often used as a starting point for other Statistics Canada surveys deemed supplementary. In this context, it is highly probable that the LFS sample allocation is not appropriate for the other surveys, but it is possible to get around this problem. Use of the LFS survey frame and sample by other surveys is discussed in Chapter 9.

The LFS sample is allocated in several steps. The first steps determine the basic allocation of the sample (described in Section 2.5.1). Section 2.5.2 describes the adjustments made to the basic allocation to better target the survey resources. The method used to implement the required sample size reduction to fund the redesign activities is presented in Section 2.5.3. Table B.4 in Appendix B provides the LFS sample allocation based on various geographical units.

2.5.1 Basic allocation

LFS quality objectives are established for the provinces, economic regions (ERs) and the Employment Insurance Economic Regions (EIERs). The purpose of the allocation is to ensure that the sample will be able to meet these objectives.

For the provinces, the CV of the monthly provincial estimate of the number of unemployed should be less than 7%. For ERs the CV of a three-month moving average should be less than 25%. As for the EIERs, the target for the estimate with a three-month moving average of the number of unemployed is 15%[7]. Since most of the census metropolitan areas (CMAs) are also EIERs, setting an objective for the EIERs also guarantees the quality of the estimates for the CMAs.

When this document was being written, two organizations were funding the LFS sample. First, Statistics Canada is funding a sample of 36,000 households to satisfy the first two objectives mentioned in the previous paragraph. The other source of funding, Human Resources and Social Development Canada (HRSDC), aims to guarantee the quality of the estimates produced for the EIERs. The sample allocation is based on the hypothesis that the Statistics Canada survey budgetary envelope is ensured over a long period of time. However, the funding from HRSDC could fluctuate over time.

In the event that the funding from HRSDC decreases, the sample funded by Statistics Canada must ensure the quality of the estimates at the national and provincial levels and per ER. To do this, these two parts of the sample must be allocated in sequence: allocation of the sample funded by Statistics Canada and allocation of the sample funded by HRSDC to improve the quality of the estimates of the number of unemployed by EIER.

2.5.1.1 Allocation of the sample funded by Statistics Canada

The first step consists of allocating the sample funded by Statistics Canada among the 10 provinces. Various strategies for carrying out this step had been considered. In the end, the redesign steering committee decided that the provincial allocation used before the redesign should be reintroduced into the new sample plan. We must mention that the strategies considered would have brought about few changes to this provincial allocation.

The second step involves allocating the provincial sample to each of its ERs. Since some ERs are small, the required sample size to produce reliable estimates would be too large. As a result, four small ERs were combined with their neighbouring ER and the quality objective was applied to these combined ERs. This situation occurs for small ERs in the northern regions of Quebec, Manitoba, Saskatchewan and British Columbia.

Allocating the provincial sample to its ERs must find a balance between the quality of the provincial estimates and the quality of the estimates for each ER. After having considered a number of alternatives, a decision was made that the provincial estimate would have precedence over the estimate by ER. In order to optimize the quality of the provincial estimates, we first allocated the provincial sample in proportion to the number of households in each ER. In this first allocation, the quality of the estimates for some ERs would not have satisfied the survey quality objectives. Therefore, minor adjustments were made.

7. These objectives are the same as in the previous redesign.

To make these adjustments, we had to predict the quality of the estimates of the number of unemployed by three-month moving average for each ER with a given sample size. This prediction was based on a CV estimation model involving the sample size, anticipated response rate, estimate of the number of unemployed, and the efficiency of the sample design for each ER. The design effect is utilized to measure the efficiency of the design used. This model is based on data from the last five years of the LFS. Therefore, we implicitly assume that the new sample design will behave similarly to the previous one.

Using the model, it is possible to quantify the impact of a change to the initial allocation on the CV of the ERs. In addition to the objectives already outlined, a constraint was imposed on each ER: we wanted at least 200 households sampled per month in each ER. The reason was to ensure that a sufficient number of households would be surveyed in each ER, and therefore to protect against errors resulting from the CV prediction model.

The sample funded by Statistics Canada is allocated to each ER by changing the initial allocation as little as possible such that: a) at least 200 households are sampled each month in the ERs; b) the monthly estimated CV of the number of unemployed by three-month moving average is less than 25%. The solution to this problem is found using a non-linear programming algorithm.

2.5.1.2 Allocation of the sample funded by HRSDC

To satisfy all three survey objectives, we must control the sample size of each ER and EIER by allocating the sample to the ER and EIER intersections. For the same reason, the survey is stratified separately at each ER and EIER intersection (see Section 2.6).

Because some ER/EIER intersections are small, it was impossible to create an effective sample design for them. Consequently, they were combined with a neighbouring intersection. One basic principle guided the combining of the small intersections: the small intersections had, inasmuch as possible, to respect the boundaries of the EIER. This approach implicitly gives more importance to the estimates by EIER than by ER because it improves the efficiency of the sample design for the EIERs. Once these intersections were combined, the entire territory of the 10 provinces was subdivided into 140 intersections.

Before allocating the sample funded by HRSDC, we needed to allocate the sample funded by Statistics Canada for one ER to the intersections within it. This first allocation is done in proportion to the size of each intersection. Following this allocation to the intersections, we derived the allocation of the sample funded by Statistics Canada in each EIER.

The allocation strategy for the sample funded by HRSDC is based on four criteria:

- The CV of the estimated number of unemployed by three-month moving average must be less than 15% for each EIER;
- The minimum sample size for each EIER is 500;
- The estimates produced by the LFS are used to establish the eligibility criteria and duration of benefits for the employment insurance program. In this context, the quality of the estimates produced for each EIER must be similar from one EIER to another.
- The portion of the sample funded by HRSDC for each province must be similar to the allocation used prior to the redesign.

This last criterion was introduced for two reasons. First, HRSDC began updating the boundaries of the EIERs during the redesign, and therefore it would have been premature to disrupt the provincial allocation of the sample funded by HRSDC while other changes will probably be required in the near future. This limitation also offers protection against potential errors in the model that predicts the efficiency of the new sample design.

Once again, non-linear programming is used to solve this problem. After allocating the sample funded by HRSDC, the sample size assigned to each EIER is finally allocated to the ER/EIER intersections. This last step is done again in proportion to the size of each intersection.

We then compared this new allocation to the one used before the redesign to identify potential errors in the model used to predict the effectiveness of the new design. Some adjustments were made following this comparison.

The allocation produces two parameters that will be used in the following steps: the inverse sampling ratio (ISR) and the number of respondents required for each intersection. The inverse sampling ratio will be used to determine the size of the strata (see Section 2.6) and used during the selection process (see Section 2.7). The

number of respondents required in each ER/EIER intersection will be used to stabilize the sample size over time.

2.5.2 Adjustments to the initial allocation

As explained at the end of Section 2.4, the collection cost in certain PSUs is sometimes very high compared to the impact of these PSUs on the estimates produced by the survey. Two methods were used to identify these PSUs: the expertise of the regional offices was used to identify an initial set of PSUs, and the vacancy rate[8] observed during the last census was used to identify other PSUs with high collection costs.

The selection of PSUs with a high vacancy rate increases collection costs because interviewers, in addition to their usual tasks, must verify whether or not the sampled dwelling is vacant for a period of six months. The collection cost for a respondent household in a PSU with a high vacancy rate is therefore abnormally high. The costs to list PSUs with a high vacancy rate are also high. From a cost-benefit perspective, it is preferable to control the same number of dwellings selected in the PSUs with a high vacancy rate.

The PSUs with a high collection cost were assigned to two groups. The first group was made up of PSUs with exorbitant collection costs. As explained above, they were identified by the regional offices. Few PSUs met the severe criteria applied to this group. To improve the return on investment of collection, these PSUs were simply excluded from the survey frame. Excluding persons belonging to the target population of a survey automatically introduces a bias into the survey estimates. However, the excluded PSUs total less than 1% of the Canadian population. Therefore, their exclusion cannot lead to a significant bias of the estimates. Table B.1 in Appendix B presents the breakdown of the number of households excluded from the survey frame for each province.

The PSUs with a high vacancy rate are assigned to the second group. To reduce collection costs, we decreased the number of households to select in these PSUs. To do this, the PSUs in the second group were assigned to special strata (see Section 2.6). This way, we can process them in a particular way without affecting the surrounding strata.

Excluding some PSUs and decreasing the sample size in the PSUs of the second group slightly disrupts the sample allocation. In fact, if nothing is done, decreasing the sample size for these PSUs will lead to a smaller sample size than that determined by the allocation. To remedy this problem, we increased the sampling ratio applicable to the PSUs with normal collection costs in the intersections containing PSUs with high collection costs.

A final adjustment was made to the sample allocation. After each census, Statistics Canada reviews the list of CMAs. After the 2001 Census, it planned to create six new CMAs for the 2006 Census. Since the LFS produces estimates for each CMA, it was important to ensure that the sample size drawn in each new CMA would be sufficient to produce good estimates. To do this, we simply took some dwellings from the sample of neighbouring regions and assigned them to one of these six CMAs. This adjustment guaranteed that we will have enough respondents to produce good estimates. However, they would be insufficient if these CMAs were promoted to the EIER level.

2.5.3 Decreasing the sample size

As explained in Section 1.4, the LFS was redesigned with limited financial resources. To complete the redesign, we had to decrease the sample size by 3% over a three-year period. The money saved during collection allowed us to fund the redesign activities.

The computer systems used by the LFS to manage the sample over time are complex. It would have been difficult to apply a specific decrease to each region and to increase the sample size in each region at the end of the three-year period. Therefore, we chose a simple option to implement the decreased sample size: decreasing it uniformly for all the regions. To apply this decrease, we used the sample size stabilization method (described in Section 3.4). The sample size targeted by stabilization was decreased by 3% without modifying the sampling ratios established during the sample allocation. At the end of the three years, we will simply have to modify the stabilization targets to restore the sample to its normal size.

Simulations indicated that the impact of decreasing the sample size on the main survey estimates will be minor. However, it could have a greater impact on the estimates of sub-populations. To minimize this risk, we plan to restore the sample to its normal size in April 2008.

2.6 Stratification

To improve the efficiency of the sample design, it is preferable to create strata. A stratum is a group of sampling units. In the LFS, the strata are sets of PSUs.

8. See the glossary in Appendix A for a definition of the vacancy rate.

Stratification involves assigning each PSU to a single stratum. Once stratification is completed, we can create a survey frame containing all of the PSUs and their corresponding strata.

The sample is selected independently in each stratum (see Section 2.7). Stratification will improve the design's efficiency if the PSUs assigned to the same stratum are homogeneous, meaning that the households therein have similar characteristics. Stratification also offers other advantages.

Since selection is done independently in each stratum, we can use a specific and more appropriate selection method for each one. Furthermore, the independence of the strata makes it possible to use a sample design adapted to the characteristics of a region. For example, for some isolated urban centres, it will be beneficial to use a three-stage sample design to reduce collection costs, while for the other regions, a two-stage design will be more effective.

Stratification also opens the door to updating the sample design for a set of strata if we observe over time that the population and its characteristics have changed greatly since the last decennial census. Lastly, we can consider operational constraints specific to certain regions to facilitate collection.

2.6.1 Changes made during this redesign

Since the 1960s, the LFS used an apartment frame, which included all the buildings in large urban centres with at least five storeys and 30 dwellings. There were two primary advantages to using it: first, it helped to better control the impact of the construction of new apartment buildings on the sample size and, as a result, minimized the impact of these buildings on the PSU selection probabilities. It was also assumed that the households residing in these buildings had different characteristics from other households.

A study of the 2001 Census revealed that the households living in apartment buildings were no different from the other households. In addition, between 1994 and 2004, the growth of the apartment frame was rather modest, which means that few new buildings were selected during this period. However, the costs associated with implementing and managing this frame were quite high. Based on this, we concluded that the benefits of the apartment frame were not great enough to justify its implementation. Therefore, we eliminated it from the sample design. In 1994, we had defined strata for low-income buildings using this frame. These strata[9] were therefore eliminated when we eliminated the apartment frame.

In the past, we used a three-stage sample design in certain rural areas with low population density. Thanks to new geographical tools, we observed that the PSUs in these strata sometimes covered a vast territory and were not always contiguous. We also noticed that the population density of these regions was not significantly lower than that of the regions for which we used a two-stage design. This is why we discarded the three-stage sample design in rural areas. Since 2004, we have been using a three-stage design only for isolated urban areas.

To reduce collection costs, we also introduced strata consisting of PSUs with a high vacancy rate. Collection in these PSUs is both costly and a source of frustration for interviewers. By isolating PSUs in specific strata, we can decrease the size of the sample selected in these areas and thus reduce collection costs. Other strata targeting rare populations were also created to meet users' needs. The methodology applied to define these strata is described in Section 2.6.3.

2.6.2 Basic stratification

Inasmuch as possible, it is preferable for the strata to respect the geographical regions for which we want to produce reliable estimates. As we saw in Section 2.5, quality objectives were defined for the ERs and EIERs. Consequently, the ER/EIER intersections form the first stratification level of the LFS survey frame. As mentioned in Section 2.5.1.2, some of these intersections are too small and must be combined with a neighbouring intersection. The combined intersections are then stratified[10] again.

It takes several steps to determine stratification. Within each of these intersections, we identify the PSUs to assign to the special strata, when necessary (see

9. We briefly considered the possibility of creating low-income household strata by grouping together PSUs with a high rate of low-income households, but this approach was discarded because these strata would have been somewhat unstable over time.

10. The Canadian Community Health Survey selects a significant portion of its sample from the area frame developed for the LFS. This survey produces estimates by health region. Therefore, we considered the possibility of taking these health regions into account to define the LFS strata. This was rejected because the boundaries of these regions change over time. Moreover, health regions intersecting with ERs and EIERs would have created many small intersections, making it impossible to create strata that respect each of these classifications.

Section 2.6.4). The remaining PSUs are then stratified geographically or optimally (see Section 2.6.5).

2.6.3 Stratum size

The LFS selection and rotation methodology adds a constraint to the stratum size. As explained earlier, we rotate one sixth of the sample every month. To implement this approach, it is preferable to select six PSUs, or sometimes 12, in each stratum. In addition, in order to improve the sample design's efficiency, studies conducted during the previous redesign revealed that it is better to select 10 households per PSU in the rural strata, eight in the urban strata, and six in the strata covering the Montréal, Toronto and Vancouver CMAs. By selecting more households per PSU in the rural strata, we hope to amortize travel costs over a larger number of units. At the other end of the spectrum, selecting six households per PSU from the three largest CMAs in the country helps to increase the number of PSUs required in the sample for these regions. Decreasing the number of households required per PSU in the three large CMAs should therefore result in a decrease in the design effect. Finally, this reduction increases the number of strata to create. More and smaller strata should lead to an increase in their homogeneity, which should also improve the efficiency of the sample design.

By combining these constraints, the number of dwellings to group together in each stratum is:

$$M_h = ISR * 6 * m_h^* \qquad (1)$$

where

M_h is the number of households to group together in each stratum of a region.

ISR is the inverse sampling ratio as established during the sample allocation. As we saw in Section 2.5, this sampling ratio is the same for all the strata in an ER/EIER intersection.

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

By dividing the number of households in a region by this result, we can determine the number of strata to create in each region. Because the result of this division is not an integer, we must round it up. We mention that with this constraint, the strata of an ER/EIER intersection are approximately all the same size.

2.6.4 Special strata

Special strata can be divided into two categories: those defined to process isolated geographical areas or areas with a low population density, and those created to target specific populations. The first category is used to determine a sample design adapted to the geography of the Canadian territory. It includes strata for remote regions, for regions with a high vacancy rate, and three-stage strata for isolated urban areas. The second category helps analysts who use LFS data to better target certain populations of interest. It includes strata with a high Aboriginal and immigrant population and a high proportion of high-income households. For simplification, we will use the terms Aboriginal strata, immigrant strata and high-income strata from now on, although this is technically incorrect since these strata do not only have Aboriginals, immigrants or high-income households.

2.6.4.1 Strata adapted to the specific characteristics of the Canadian territory

A significant part of the territory in Canada is inhabited by a small portion of the population. Collection costs in regions with a low population are very high, while the impact of these regions on the main LFS estimates is relatively low. Therefore, it is necessary to develop approaches suitably adapted to these regions in order to effectively assign our limited resources.

To satisfy this objective, we begin by defining strata for remote regions. These strata include the parts of Canada with the lowest population density. Once these regions are assigned to specific strata, we can better control the size of the sample selected in these strata, and thereby better control the assignment of our resources. The boundaries of these regions are essentially the same as in the previous redesign, but some adjustments were made to exclude the regions whose population density has increased since 1991. Then, a stratum for remote regions was developed for the northern part of all the provinces, except Prince Edward Island, Nova Scotia and New Brunswick. These strata consist of contiguous territories.

As we saw earlier, collection costs are abnormally high when the vacancy rate in a PSU is high. The population density of these PSUs is also low, and they are sometimes located near an urban centre. This characteristic sets them apart from PSUs in remote regions. Strata with a high vacancy rate include these

PSUs. Contrary to the strata of remote regions, strata with a high vacancy rate do not cover a contiguous territory.

The last stratum type in this category provides a solution to the operational limitations associated with collection in these average-sized, geographically isolated urban centres. These centres are too small for interviewers to complete their task over a long period. However, they are too isolated from other urban centres for an interviewer residing outside its borders to be able to travel there for collection. To reduce costs associated with travel and training to hire a new interviewer, it is preferable to use a three-stage sample design to cover these urban centres.

In the first stage, we select an urban centre from all the centres in a stratum[11]. The territory of the selected urban centre is divided into plots (secondary sampling units (SSUs)). The approach used to define the PSUs is also used to define the SSUs. All the SSUs of a PSU are assigned to the same rotation group. Therefore, an interviewer can contact all the new households introduced into the PSU sample in the same month. Since the first interview in a household is often done in person, this approach should reduce travel for the interviewer. In the second stage, we select an SSU sample. In the third and final stage, we draw a sample of dwellings in the selected SSUs.

This approach guarantees that a sampled urban centre will remain in the sample for a long time. It also reduces collection costs by concentrating the sample in fewer of these small urban centres. Three-stage strata were created in Quebec, Ontario, Alberta and British Colombia.

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

2.6.4.2 Strata to target certain sub-populations

The sub-populations targeted by these special strata are relatively rare. An area frame is not a very effective tool to target sub-populations when these are not concentrated geographically.

The members of those sub-populations of interest to us do not all live in the same neighbourhoods. Also, in a neighbourhood with several members of a given sub-population, many households have no members in this sub-population. As a result, the prevalence of these sub-populations in the special strata is relatively low. It is important to note, however, that this prevalence is much greater than the one observed in the entire Canadian territory. Within the scope of the sample design used by the LFS, they are, although not perfect, the only tool available to target these sub-populations.

Three criteria determine the effectiveness of the special strata. As mentioned in the previous paragraph, the first is prevalence. The second is the proportion of the target population residing in the PSUs assigned to the special strata. These two criteria go hand in hand: they succeed or fail together. For example, an Aboriginal stratum with a 60% prevalence of Aboriginal households would not be very effective if it only covered one percent of the Aboriginal population. The opposite is also true. The third criterion is the stability of these strata over time. The initial implementation of high-quality special strata is futile if they deteriorate very quickly after a few months.

One final crucial factor was considered for developing the special strata creation strategy: their impact on the estimates produced for the total population. The result of good 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, we used 2001 Census data to measure the impact of various scenarios on prevalence, the proportion of the target population covered by the special strata, and the impact on the main survey estimates. The 1996 Census data were used to evaluate the stability of each scenario over time. The guidelines for the creation of special strata are based on this study.

The first guideline states that the strata must be created based on the prevalence of specific characteristics. For example, given that the proportion of immigrants in Prince Edward Island is very low, it would be futile to create an immigrant stratum for this province. Based on this criterion, Aboriginal strata can only be created in Manitoba, Saskatchewan, Alberta and British Columbia. Likewise, immigrant strata are required for the Montréal, Ottawa, Toronto, Calgary and Vancouver CMAs. Finally, we can create high-income strata in the country's biggest CMAs.

The second guideline specifies a limit to the number of special strata that can be created. This limitation guarantees that these strata will not have a major adverse effect on the main LFS estimates. The study conducted using the 1996 and 2001 Census data illustrated that each special strata category should not cover more than 8% of the population of a region.

Based on these guidelines, the special strata were created in sequence. For each category, they were

11. These urban centres correspond to the boundaries of the census subdivisions.

created by identifying the PSUs with the highest prevalence of the sub-population of interest. Using this approach, these strata are not contiguous.

We began by creating the high-income strata in the biggest CMAs. To do this, the PSUs in a given CMA were first classified in descending order based on the proportion of households with an income over $125,000 based on the 2001 Census[12]. The PSUs at the top of this list were assigned to a high-income stratum until its predetermined size had been reached (see Section 2.6.3). If the limit of 8% was not attained, another high-income stratum was created for the same CMA. Once the work for this CMA was completed, we moved on to the next CMA.

The same approach was applied to the immigrant and Aboriginal strata. To create the immigrant strata, the PSUs were put into descending order based on the proportion of households with at least one immigrant based on the 2001 Census. In Montréal, Toronto and Vancouver, a household was classified as immigrant if at least one member had immigrated to Canada in the previous five years. The rule for Ottawa and Calgary was based on entry into the country in the previous ten years, since there are fewer immigrants in these CMAs.

To create the Aboriginal strata, we had to modify the basic strategy slightly. The high-income and immigrant strata respect the CMA boundaries, but a significant number of Aboriginals live outside these boundaries. Furthermore, some ER/EIER intersections 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 respect the boundaries of the EIERs, rather than those of the ER/EIER intersections. Finally, the PSUs already assigned to a remote region stratum, a high vacancy rate stratum or a three-stage stratum could not be assigned to an Aboriginal stratum[13].

Table B.3 in Appendix B gives 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.

12. We considered the possibility of measuring prevalence using tax data rather than the 2001 Census. However, we opted to use the census because it was impossible to correctly assign each tax record to a PSU.

13. This limitation is not required for the immigrant and high-income strata because the CMAs are not affected by these special strata.

2.6.5 Stratification of the remaining PSUs

The strategy used to define several types of special strata was described in the above section. However, these strata only cover a small portion of the Canadian territory. This section describes the approach used to stratify the normal PSUs, or those not assigned to a special stratum.

Over time, the demand for information on census subdivisions (CSDs) has increased sharply. As a result, we have tried to create strata that respect the boundaries of the CSDs. With this approach, we stabilize the size of the sample selected in the CSD and thus improve the quality of its estimates. Since several CSDs are relatively small, it is impossible to create a stratum therein because the LFS imposes a limitation on the number of households that form a stratum. Equation (1) in Section 2.6.3 gives the target size for each stratum. It also specifies the minimum size of a CSD in order to create one or more strata that respect its boundaries. As a general rule, we were able to create strata in CSDs with at least 20,000 households.

Where possible, we also tried to create strata that respect the urban and rural areas. This approach is justified for three reasons: rural strata have more households than urban strata (see Equation (1) in Section 2.6.3); persons residing in rural areas have different characteristics from those residing in urban areas; and, stratification that respects these areas allows us to implement more appropriate collection strategies.

These three basic rules serve as a framework for stratifying the remaining PSUs. The stratification strategy operationalizes these rules into a sequence of steps to carry out. In each step, we must determine the number of strata to create and decide how to process the small areas created from the intersection of the boundaries of the CMAs, ERs, EIERs, and urban or rural areas.

Since we want to produce estimates of good quality for the CMAs, it is preferable for stratum boundaries to respect CMA boundaries. We saw earlier that the CMA boundaries generally respect those of the EIERs. On rare occasions, these boundaries are not exactly the same. For these CMAs, it is impossible to create strata that respect both the CMAs and the ER/EIER intersections. A decision was made in each case to minimize the impact of the slight adjustment to one of these boundaries on the CMA and ER/EIER estimates in question. The same situation arises when we want to create strata that respect the urban and rural areas. In some cases, the urban area is too small to create a stratum within it, in which case it is necessary to

combine these areas with a neighbouring urban area or a rural area. Once again, each case was evaluated separately. The small rural areas are treated the same way.

After resolving these boundary problems, we stratify each of the CMAs separately. In a given CMA, we first identify the CSDs large enough to form strata that respect their boundaries with the help of Equation (1) in Section 2.6.3. We then check whether it is necessary to create more than one stratum by dividing the number of households in the CSD by the targeted stratum size. Since this quotient is not an integer, we round it up. If the CSDs are large enough to create at least 10 strata, we first create superstrata, which divide the CSD territory into compact areas made up of a similar number of households. This approach ensures better geographical distribution of the selected PSUs. Finally, we perform optimal stratification in each CSD with more than one stratum and in each superstratum.

The purpose of this stratification is to reduce the sampling variance of several variables of interest by grouping together PSUs with similar characteristics. The list of these variables of interest (29 in total) is identical to the list from the last redesign and is available at the end of Appendix B.

The algorithm implemented to do this stratification is based on a method developed by Friedman and Rubin (1967) and modified by Drew, Bélanger and Foy (1985). The purpose of these modifications was to adapt it to the context of surveys with unequal probabilities and to produce strata of similar size. The algorithm helps to give certain variables more importance. For our purposes, the same importance is given to each one, except for household income, which is three times as important, because income is correlated to several variables.

The algorithm uses an iterative approach. Using an initial stratification that respects the limitations, it exchanges a PSU between two strata and checks whether this new stratification decreases the variance. If so, this new solution replaces the previous one. If, however, it increases the variance, the previous stratification is retained. This exchange process is repeated until no exchange leads to a decrease in variance. Once completed, we use other different initial stratifications and repeat the process. The stratification associated with the smallest variance among all the variances considered is retained. It is possible to create compact strata that are contiguous with this method. These additional constraints greatly limit the PSU combinations that may be considered; therefore they produce solutions with greater sampling variance. This is why the strata created during optimal stratification are neither contiguous nor compact, as they were in the last redesign.

Once the large CSDs are stratified, we move on to processing other urban areas. The strategy used for the large CSDs is used once again. In each urban area, we begin by determining the number of strata required. If more than 10 strata are required, we create superstrata. We then carry out optimal stratification by urban area or by superstratum to create the final strata. The same steps are carried out for the rural areas.

This approach is also used to stratify the PSUs outside CMAs. In each ER/EIER intersection, we begin by identifying the CSDs large enough to contain at least one stratum. The rest of the territory is then stratified optimally based on belonging to the urban or rural area. We must mention that the urban area outside CMAs is fragmented, meaning it is made up of small cities spread out over the territory of the intersection and surrounded by the rural area. As a result, the PSUs in the same strata can be located several kilometres from one another. This fragmentation does not result in increased collection costs because we use the same sampling rate for the rural strata surrounding the urban strata. In this context, a PSU selected in an urban stratum will often be close to another selected PSU. As explained earlier, the geographically isolated urban centres are included in the stratum with three stages of selection.

2.7 Sample selection

Once stratification is completed, all the parts are in place to select the sample. This section provides a conceptual description of the selection and rotation method used by the LFS. Additional information on processing the growth and maintenance of the survey frame is given in Chapter 3.

When we use a two-stage design, 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 condition is satisfied 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 selected with a probability proportional to their size. The size measurement we use to calculate the selection probability of each PSU is the number of households in the PSU based on the 2001 Census[14].

14. In practice, we actually derive a size measurement from the number of households. This measurement is called the inverse sampling ratio for each PSU. More information on this calculation is provided in Section 2.7.1.

The sample is selected independently in each stratum, which makes it possible to adopt a different selection method for each one. There are several methods for selecting a PSU sample with probability proportional to size. The LFS uses two of these methods, one of which is applied to most strata. These two methods are presented in Section 2.7.1.

As explained in Section 2.1, we replace 1/6 of the sample every month. To simplify the sample rotation process, it is preferable to select six PSUs, or a multiple of six, in each stratum. This way, we simply replace the dwellings selected in one of these six PSUs every month to do the rotation. We will come back to this rotation method in Section 2.7.2.

To determine the number of PSUs to select in the stratum, we first calculate the number of households to survey in the stratum based on the sample allocation. To do this, we divide the number of households in the stratum by the inverse sampling ratio (ISR) established during the sample allocation. We then divide this quotient by the ideal number of households to survey per selected PSU. As we saw in Section 2.6.3, this ideal number is six per PSU in the Montréal, Toronto and Vancouver CMA strata, eight in the urban strata outside these three CMAs, and 10 in the rural strata. If the result of the second division is closer to six than to 12, we will select six PSUs in the stratum. Otherwise, we select 12. When stratifying, the stratum size is determined based on the hypothesis that six PSUs would be selected (see Section 2.6.3). Consequently, the second division will lead us to the conclusion that we must select six PSUs in the large majority of strata. Explanations on selecting six PSUs are provided below. The same approach applies when we want to select 12 PSUs.

2.7.1 Selecting PSUs and the first sample of dwellings

The LFS uses two methods to select the PSU sample: the Rao-Hartley-Cochran (RHC) method, and the systematic sampling method with probability proportional to size and random order. The RHC method is used with most strata, because it allows us to update the selection probabilities when we observe strong growth 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.

We present a summary of the RHC method below. For more information, see Rao, Hartley and Cochran (1962). Because the LFS seldom uses the randomized probability proportional to size systematic sampling method (RPPSS), we do not provide a description of it here. The principles described for the RHC method also apply to the systematic sampling method. If you would like more information on this method, please consult Cochran (1977).

To select six PSUs in a stratum, we must first distribute all the PSUs within the stratum into six groups containing the same number of PSUs. Each group will be associated with a rotation group (see Section 2.1). Then, we simply select one PSU per group with probability proportional to size in the group. This can be summarized by the following equation:

$$\pi_{hij} = \frac{M_{hij}}{\sum_{j \in hi} M_{hij}} \qquad (2)$$

where

M_{hij} is the number of households in PSU j in group i of stratum h based on the 2001 Census.

$\sum_{j \in hi} M_{hij}$ is the sum over all the PSUs in group i of stratum h.

π_{hij} is the selection probability of PSU j in group i of stratum h.

In the second stage, we want to draw a sample of households in the selected PSUs. To simplify this process, the LFS selects the households using systematic sampling. This method is recommended because it is simple to use, ensures a good distribution of the households selected in the PSU and facilitates adding new dwellings to the PSU. To make a selection using systematic sampling, we must determine the sampling interval, which is the inverse sampling ratio (ISR) of the PSU. It is established using the number of households in the PSU based on the 2001 Census and the ISR determined during allocation of the sample. It is calculated using the following equation:

$$ISR_{hij} = \left(\frac{M_{hij}}{\sum_{j \in hi} M_{hij}} \right) ISR_h \qquad (3)$$

where

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

ISR_h is the inverse sampling ratio of stratum h established during allocation of the sample[15].

15. As explained in Section 2.5, we use the same ISR for all the strata in an ER/EIER intersection.

Since ISR_h is constant for all the PSUs of a group, ISR_{hij} is proportional to the number of households in each PSU. Consequently, it is possible to select a sample with probability proportional to size using these ISR_{hij} as size measures. The LFS selection system is configured to use integer inverse sampling ratios. The result of Equation (3) is therefore rounded up or down so that $\sum_{j\in hi} ISR_{hij} = ISR_{hi} = ISR_h, \forall i \in h$.

Afterward, there are two more intuitive interpretations for ISR_{hij}. According to the first, ISR_{hij} is the number of distinct samples available in the PSU. In LFS jargon, this concept is called the number of seed values. The PSU inverse sampling ratio is also the sampling interval to use if the corresponding PSU is selected in the first stage. By applying this sampling interval, we will select the appropriate number of households in the PSU and attain the target for the group[16].

In short, Equation (2), following controlled rounding, provides the size measure to use when selecting a sample with probability proportional to size. The first-stage selection probability associated with each PSU is therefore:

$$\pi^*_{hij} = \frac{ISR_{hij}}{\sum_{j\in hi} ISR_{hij}} = \frac{ISR_{hij}}{ISR_h}. \qquad (4)$$

Furthermore, the second-stage selection probability when PSU *j* in group *i* of stratum *h* is selected is $1/ISR_{hij}$. As a result, the selection probability of household k in PSU *j* in group *i* of stratum *h* is:

$$\pi^*_{hij} = \frac{ISR_{hij}}{\sum_{j\in hi} ISR_{hij}} \times \frac{1}{ISR_{hij}} = \frac{1}{ISR_h}. \qquad (5)$$

The selection probability of all households in the same stratum is the same. The LFS sample design is therefore self-weighted[17]. Finally, we will note that $\pi^*_{hij} \approx \pi_{hij}$ because

$$\pi^*_{hij} = \frac{ISR_{hij}}{\sum_{j\in hi} ISR_{hij}} \approx \frac{M_{hij}}{\sum_{j\in hi} M_{hij}} = \pi_{hij}. \qquad (6)$$

The difference between these two probabilities is due to the rounding of ISR_{hij}.

In practice, to select a PSU in a group, we begin by putting the PSUs of a group in random order. We then draw a random whole number U from a uniform distribution on the interval $U \in [1, ISR_h]$. 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 equal to or less than *U* (or $\sum_{j\le k} ISR_{hiij} \le U$ where the indicator j follows the random order).

It also determines the number of seed values to use in this PSU *k* before moving on to the next PSU. The number of values to use in the first PSU is $D_k = (\sum_{j\le k} ISR_{hij}) - U + 1$. Lastly, we select a second random whole number $U_k \in [1, ISR_{hik}]$[18], which indicates the first seed value to use to select the sample of dwellings for the PSU *k*. These dwellings will remain in the sample for a period of six months.

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.

2.7.2 Sample rotation

Section 2.7.1 describes how we select the first sample of dwellings 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 seed value U_k of the PSU *k*.

If the number of seed values to use from PSU *k* is $D_k = 1$, the second sample of dwellings will correspond to the value U_{k+1} of PSU $k+1$ where $U_{k+1} \in [1, ISR_{hi(k+1)}]$. Otherwise, the second sample will correspond to the value $U_k + 1$ of PSU *k*. If $U_k + 1 > ISR_k$, we move to value 1 of PSU *k*. Generally speaking, with this method, PSU *k* remains in the sample for D_k periods of six months. When it is necessary to replace the surveyed dwellings, we simply move to the next seed value. After D_k periods, we move to the value U_{k+1} of PSU $k+1$. This PSU will remain in the sample until all its seed values have been used. The same goes for the PSUs that are added to the sample at a later date.

16. This target corresponds to the number of households in the group (based on the 2001 Census), divided by the stratum sampling ratio.

17. A sample plan is self-weighted if all the units in a region have the same weight. For the LFS, all the units in the same ER/EIER intersection have the same survey weight.

18. This second random number has two functions. It takes into account the fact that the sample size associated with the last seed values is sometimes smaller than that of the first values. 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 values to use.

This method produces the expected results: the selection probabilities are always respected over time. Unfortunately, it has a major inconvenience. As we saw, the first PSU remains in the sample for a random number of periods. In some cases, the first PSU selected remains in the sample for few periods. This rapid rotation of the first PSU selected would to lead to an inefficient use of our limited resources. In fact, adding a PSU to a sample requires a great deal of work on our part, including preparing the material, listing it 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 seed values to use from the first PSU without introducing a bias into the selection probabilities. When D_k is too small, based on a pre-determined criterion, we increase it in order to keep this PSU in the sample longer. In this case, we must proportionally reduce the number of values to use for PSU $k+1$ in order to avoid introducing a bias into the selection probabilities. Some constraints are required to ensure that the increase in the number of values associated with the first PSU will not reduce the number of values 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 group created using the RHC method. 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, we can determine, the day after the redesign, the list of values that will be in the LFS sample for each month over the next 10 years.

It is important to note that the rotation method described in this section, including the adjustment made to the number of values associated with the first PSU selected in the group, also applies when the sample is selected using the systematic selection method with probability proportional to size and random order. Gray (1973) provides proof of this.

Chapter 3 Sampling frame creation and maintenance

3.1 Use of the Address Register

3.1.1 Introduction

As described in the previous chapter, the LFS has a two-stage sample design with an area frame at the first stage. Thus, a complete list of dwellings within each selected Primary Sampling Unit (PSU) is needed to select the second stage sample. Historically, the list of dwelling addresses used to form the second stage sampling frame was obtained through a listing exercise performed in the field for each selected PSU. To avoid this expensive exercise, a new source of addresses is now being used: the Address Register (AR), a list of nearly 90% of the dwellings in Canada. The AR had the potential to reduce the cost and time associated with the creation of the second stage frame, and possibly even increase its quality.

The availability of a list of addresses opened the door for the development of a simpler, more efficient one-stage design rather than the two-stage design used for the LFS in the past. However, because of time and budget constraints, it was decided that the two-stage design would be kept for now. Therefore, a strategy to use the AR in the best possible way in the context of a two-stage design was developed and implemented, as described in this chapter.

3.1.2 The Address Register

The AR 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. After the Census, it was updated using the list of addresses created during the Census enumeration process. Since then, the process has remained the same: the AR gets updated through administrative files before the Census takes place and is further updated using the information later gathered by the Census. For a more detailed description of the creation and maintenance of the AR see Swain, Drew, Lafrance and Lance (1992).

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 the 2001 Census to include less-populated regions. Because of the history of maintenance for large communities, the AR is more accurate in these communities. With time, it should become more accurate in less-populated regions as well.

Since the Census is only carried out every five years, the accuracy of the AR can deteriorate between Census updates. While most of the changes will be due to new dwelling construction, other changes are due to businesses being converted to private dwellings (and vice versa), single family dwellings being converted into apartments, dwellings being demolished, *etc*.

In spite of these two coverage issues, the overall coverage rate of the AR was estimated from post-Census studies to be approximately 96% in the covered areas. On the other hand, the coverage rate was also known to vary from region to region, an important factor to consider when developing a methodology for using the AR in the context of the LFS.

In 2005, the AR included approximately 13 million addresses. The majority of these addresses were reported to be valid residential dwellings during the 2001 Census. Some of these addresses, which predate the 2001 Census, were retained even though they were found to be invalid following the 2001 Census. Other addresses found on the AR are obtained through updates from administrative files (in preparation for Census 2006).

To appear on the AR, a residential dwelling must possess a valid standard civic address. Hence, we can expect undercoverage in rural areas where some residential dwellings do not have a valid civic address.

Collective dwellings are another category of dwellings available through the AR extraction process; however they do not reside on the AR. A complete list of collective dwellings is created during the Census and is used to create the Collective Dwellings List Frame (CDLF). This frame is not updated through the regular AR updating process, and remains relatively static between Censuses. Collective dwellings are part of the LFS target population and therefore should be covered by the sampling frame

When a list of residential addresses is created through the AR for a particular PSU, the collective dwellings that appear on the CDLF and that can be associated with the corresponding AR area are added to the list. The definition and treatment of collective dwellings is discussed in more detail in Section 3.3.

3.1.3 The AR and the National Geographic Database

In order to use the AR in a two-stage design context (as described later in Section 3.1.5), we must be able to assign addresses to a specific PSU. There are two ways

to do this. The first is by linking the AR to the National Geographic Database (NGD), which contains the street network of the entire country, along with address ranges for most street sections, also called blockfaces. Like the AR, the accuracy of the NGD can vary from region to region. For all addresses linked to the NGD, through the address ranges, we know which blockface they fall on, making it possible to assign them to a specific PSU. Such addresses are called 'structured'. For addresses that do not link to the NGD but were valid during the last Census, we can use another geographical link called the Census block, which is the city block that the addresses belonged to in 2001, according to the Census enumerator. Since the PSUs were built using Census blocks as a base (as explained in the previous chapter), we can also assign those addresses to a specific PSU. Since we do not know exactly which blockface they fall on, we call these records 'unstructured'. At the time of the redesign, about 86% of all the addresses on the AR were 'structured'.

The 2001 city block is available for all valid records of 2001, but, to be consistent with the PSU map that is prepared from the NGD, it was decided to use the block-face link first, when available. It is important to note that any growth record, as well as any pre-2001 address found to be invalid in 2001, that cannot be blockface geocoded, cannot be assigned to a specific PSU (we do not have a 2001 city block for them) and is consequently 'lost' for the LFS[19].

3.1.4 The sequencing algorithm

For each selected PSU, a list of addresses is prepared and sent into the field for verification. The addresses need to be put in a specific order to facilitate and optimize the listing exercise. An algorithm was developed which lists the blockfaces 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. Since the sequence is defined from the blockfaces, it is impossible to use this algorithm to position the unstructured records in the right place on the list. Therefore, they are added to the end of the list, sorted by street name and civic number. For more details on the sequencing algorithm, as well as a discussion of its strengths and weaknesses, see Laflamme and Dochitoiu (2005).

19. A new step is currently being implemented to assign a specific PSU to a growth record that is not blockface geocoded but that appears on a street that is entirely containred with the boundaries of the PSU.

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 quality of the CDLF.

3.1.5 Use of the AR

As mentioned earlier, because of time and budget constraints, the possibility of using the AR as a list frame in a one-stage design was put aside early in the development stage. Therefore, ways to incorporate its use within the traditional LFS two-stage design were studied instead. A few possibilities were considered, but in the end, the most promising one was the following.

1. Based on the estimated AR quality within each PSU, the PSUs are divided into three groups:
   a. In AR Group 1, where the AR is expected to be of excellent quality, no pre-listing in the field is performed and the initial sample of dwellings is directly selected from the AR-based list for a PSU.
   b. In AR Group 2, where the AR is expected to be of good quality, a preliminary list is created from the AR and then verified and updated by the field interviewer during a listing exercise. The route determined by the sequencing algorithm appears on the interviewer's map and should be followed.
   c. In AR Group 3, where the AR is of poor quality or nonexistent, no use of the AR is attempted and a traditional field listing is performed. The route determined by the sequencing algorithm appears on the map and should be followed, as in AR Group 2.
2. The list of addresses in each PSU, no matter its AR Group, is maintained (*i.e.*, updated) in the field at least once a year, as was done in the past. In the future, if the AR starts being updated sub-annually, it will be possible to replace this field exercise in most cases.

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. This strategy also requires that a measure of the AR quality at the PSU level be developed, which is not a simple task.

A field test was performed in the fall of 2003 to test the proposed strategy and to help develop rules for assigning PSUs to AR Groups. This test showed that, in terms of overall undercoverage, using an unverified AR as a frame (as in AR Group 1) would give results similar

to the traditional method of listing. On the other hand, the overall overcoverage would be much higher. Fortunately, after verifying the AR list (as in AR Group 2), the undercoverage becomes less of an issue and is even better than the undercoverage under the traditional method. As well, the overcoverage decreases substantially, although it seems to stay somewhat higher than the overcoverage obtained when using the traditional method. In light of these results, the proposed strategy was accepted and implemented as described in the next section. To learn more about the field test, see Thompson and Turmelle (2004).

Before the AR extractions can be used in the field for listing purposes (AR Group 2) or for sample selection (AR Group 1), a certain amount of data processing is needed. This consists of eliminating duplicate records, transferring some addresses from one PSU to another, removing units within collective dwellings and reformatting the lists to make them compatible with LFS systems. See Gouzi *et al*. (2004) for a complete description of the editing and manual intervention performed on the AR.

3.1.6 AR Group allocation strategy

The main challenge for the implementation of the AR Group strategy was to develop a method to properly assign PSUs to the three AR Groups. There were two competing goals: reducing the number of PSUs requiring field verification (or maximizing AR Group 1) and maximizing the quality of the resulting frame. The quality of the frame will be affected by the quality of the AR for the PSU, the quality of the listing sequence and by the performance of the interviewers who verify the PSU. One recommendation made by the interviewers who participated in the field test was to minimize the number of unstructured addresses on the lists sent for verification. Unstructured addresses are not sequenced properly (they appear at the end of the list), so having a lot of them on the list can involve a lot of driving and/or walking to verify their status, making the listing exercise quite tedious and prone to error. All of these concerns were taken into account when developing the strategy.

The initial PSU allocation strategy was developed using the results from the field test as well as the results from the 2004 Census test. From these two tests, a set of important characteristics was identified and then used to develop a series of rules for assigning PSUs to the AR Groups. The main characteristics used were:

- The number of unstructured dwellings in the PSU and how scattered they are (one hundred single family houses on 15 different streets is more problematic than one big apartment building).
- The AR coverage after the 2001 Census.
- The number of Census blocks in the PSU (since the sequencing algorithm does not always perform well when there are many Census blocks).
- The number of growth records added since the last Census.
- The number of multi-unit buildings of size two or three (this category of dwellings is known to contain a fair amount of overcoverage and therefore it should be verified).

The criteria used to classify a PSU in one of the three groups are:

- AR Group 3

  If (a) there are too many unstructured dwellings or they are too scattered or (b) the AR coverage in 2001 was less than 90% or (c) the number of Census blocks in the PSU was greater than 20, then put the PSU into AR Group 3. It was deemed inappropriate to use the AR for these PSUs.

- AR Group 1

  Otherwise, if (a) the AR coverage in 2001 was between 97.75% and 103% and (b) the number of unstructured records is very low or very concentrated and (c) the amount of size two or three multi-unit buildings is small enough and (d) the number of records added since the last Census is very low and (e) the number of collective dwellings is less than five (with a maximum of one unstructured collective dwelling), then put the PSU into AR Group 1. The idea was to put only the very low risk, stable PSUs into Group 1.

- AR Group 2

  Otherwise, put the PSU into AR Group 2.

These rules were applied to almost all PSUs. Manual intervention was needed for only a few problematic PSUs. In the end, the initial allocation assigned about 39% of the sampled PSUs to AR Group 1, 24% to AR Group 2 and 37% to AR Group 3. This is the allocation that was used in production when the listing exercises started at the end of the summer in 2004.

After a couple of weeks of listing, we received some feedback from the field that was quite negative. One common complaint was that sometimes the street network shown on the PSU map was out-of-date, especially in high-growth areas. Another complaint was related to the sequencing algorithm: sometimes the route did not make sense to the interviewer and/or seemed far from optimal. These two issues made the listing exercise confusing and burdensome, which consequently increased the time and money spent listing problematic PSUs and also increased field staff's frustration towards the new methodology.

Since nothing could be done at that point to improve the quality of the street network or the optimality of the sequencing, the problem was addressed by modifying the way things were done in the field. For extremely problematic cases, it was decided to let interviewers do the listing in the traditional way, which meant:

- For AR Group 3 PSUs: ignore the pre-determined route that appears on the map and use experience and judgment to come up with the optimal route.
- For AR Group 2 PSUs: ignore the AR list and the predetermined route and start from scratch as for the problematic AR Group 3 PSUs. Ignoring the AR list is simple in Initial Listing since all dwellings on the list can simply be deleted.

Since the new LFS sample was phased-in one rotation group at a time over a period of six months, the listing exercise was also done one rotation group at a time. This gave us the opportunity to revise the rules used to assign PSUs to AR Groups, based on the new information from the first rotation group, for the other rotation groups. We meticulously examined several problematic PSUs that had already been listed and identified some common characteristics that had not been used so far and that could help in identifying problematic PSUs. The main characteristics that seemed to be linked to problematic PSUs were: a large increase (or decrease) in the length of the street network since the last Census, and the addition of new streets since the last Census. We used these characteristics to adjust the rules, and we also tightened some of the preliminary rules. In the end, we transferred over three hundred sampled PSUs from AR Group 1 and 2 to AR Group 3. A complete description of the development process for the rules, as well as a description of the manual work that had to be done to assign some PSUs is given in Thompson and Rodrigue (2005).

The final allocation of PSUs to AR Groups, at the population level and at the sample level, is given in Tables 3.1 and 3.2.

The distribution among AR Groups varies greatly from province to province. The AR is not as useful in the Atlantic provinces (the first four rows), but is much better in the other provinces, especially in Ontario, Manitoba and British Columbia, where over 40% of the sampled PSUs are in AR Group 1.

This distribution will be dynamic over time. As PSUs rotate in and out of the sample and as the quality of the AR evolves (especially after the 2006 Census) the distribution of the AR Groups will likely change.

Table 3.1 Distribution of PSUs into Address Register Groups

Province	Group 1		Group 2		Group 3		Total
	number	%	number	%	number	%	number
Newfoundland and Labrador	100	10	164	17	729	73	993
Prince Edward Island	48	17	45	16	193	67	286
Nova Scotia	357	21	324	19	1,002	60	1,683
New Brunswick	252	18	200	14	964	68	1,416
Quebec	5,824	40	3,366	23	5,465	37	14,655
Ontario	9,276	46	4,560	23	6,131	31	19,967
Manitoba	1,108	52	237	11	771	37	2,116
Saskatchewan	618	31	269	13	1,121	56	2,008
Alberta	2,250	42	1,129	21	1,973	37	5,352
British Columbia	2,945	40	2,269	31	2,122	29	7,336
Total	22,778	41	12,563	22	20,471	37	55,812

Note: See Appendix A.2 for abbreviations.

Table 3.2 Distribution of sampled PSUs into Address Register Groups

Province	Group 1		Group 2		Group 3		Total
	number	%	number	%	number	%	number
Newfoundland and Labrador	27	12	38	17	157	71	222
Prince Edward Island	25	20	21	17	80	63	126
Nova Scotia	88	26	59	17	197	57	344
New Brunswick	58	18	45	14	217	68	320
Quebec	429	33	314	23	553	43	1,296
Ontario	935	46	459	22	644	32	2,038
Manitoba	215	51	49	11	162	38	426
Saskatchewan	151	35	80	18	207	47	438
Alberta	213	38	125	23	220	39	558
British Columbia	333	41	242	30	232	29	807
Total	2,474	38	1,432	22	2,669	40	6,575

Note: See Appendix A.2 for abbreviations.

3.2 Mapping

3.2.1 Introduction

The LFS Design takes advantage of the fact that, since 2001, the entire country is available in the framework of a Geographical Information System (GIS). The 10 provinces were divided into PSUs using a modified Geographic Area Delineation System (GArDS) as described in Section 2.4. Once the PSUs were delineated, including any manual interventions, a number of new features were incorporated to produce PSU maps.

- The boundaries of our PSUs were added to Geography Division's warehouse of geographical data, making LFS boundaries accessible to GIS software applications.
- Geography Division produced automated map-sizes and inset tables sufficient to map each PSU and any small details within as explained in the next section.
- Address Register (AR), in conjunction with Geography produced a sequence of blocks, blockfaces within blocks and addresses within blockfaces for any PSU covered by the AR.
- The sequence of block numbers and the optimal starting point for listing within each block were added to the map layers.
- Geography developed a mapping application to create, browse and print PSU maps.

Stratification and initial PSU selection generated over 6,500 PSUs for the introduction of this design. Once the design sample is introduced, PSU replacement generates an additional 30-60 PSUs per month. The GIS software, Arc/Info, can be used to browse, print and output maps in a wide variety of formats and styles. These maps are called *F01 Cluster Diagrams* in the field, and are used by the interviewers to list dwellings within PSUs.

The steps to identify, map, print, verify, update and ship these PSUs to the Regional Offices are detailed below.

3.2.2 LFS Mapper

Geography Division created an Arc/Info application called LFS Mapper to generate *F01 Cluster Diagrams* in Arc/Info MXD format and Adobe PDF format for every selected PSU. Arc/Info MXD files allow for the manual manipulation of individual PSUs as the need arises. Adobe PDF outputs are used for printing and display purposes. The LFS Mapper software uses geographical information including the road network, Census geography layers, various physical features and the LFS PSU boundary layer, that are stored in a central server environment within Geography division. For additional details see Cillis (2004a).

The main map for each PSU is produced in one of three sizes: 11"x17", 17"x22" or 22"x34". For any areas that do not display sufficient detail on the main map, insets are generated with a map size of 11"x17". Most PSUs have no insets, but some rural PSUs covering large tracts of land may have 10 or more insets. Naturally there is some interplay between the number of insets and the map size; large maps require fewer insets. On average each PSU requires almost 0.6 insets turning the 6,500 PSUs into more than 10,000 maps.

A database of PSUs, map sizes and inset coordinates determines the set of maps to generate for each active PSU. The PSU table contains all the sampling parameters to be printed in the legend of the map. A flag can be set to allow batch processing of an entire month's set of PSUs. The PSU table is updated monthly with each new set of PSU replacements.

A menu driven user interface permits a variety of functions as detailed in Cillis (2004b). In summary:

- the MXD files can be edited,
- the PDF can be printed,
- the table entries for a PSU can be modified for a given map size and inset deletion or creation, and
- the creation of new MXD and PDF files can be controlled by date of introduction for new PSUs or individually by PSU identification.

Additional features can be added to the LFS Mapper, including a flag to automatically turn off starting points and block numbers if the sequencing of blocks is not required, and the use of colours instead of grey-scale.

3.2.3 Sample Design System

To complement the LFS Mapper software, enhancements were made to the Sample Design System. A PSU-map control system manages the thousands of maps to be processed. An automated print system allows for batch processing of large volumes of PDF printouts to multiple printers. An interface system transfers listing data from the Address Register extracts to the Sample Listing System. A working copy of every map is generated for use by the AR data quality verification described in Section 3.1.

3.2.4 Map verification

Clerical verification of the *F01 Cluster Diagram* is required to adjust the map size, delete or create insets, add missing street names, indicate civic numbers where appropriate, add in descriptions where large PSUs are split, remove block numbers and starting points where appropriate and add notes from the AR data quality verification.

Map sizes and inset boundaries are automatically generated but may not be appropriate in all cases. Many collectives and apartment buildings are given a sub-block boundary that usually generates an inset. Sub-blocks are an artifact of the Census block program, which attempts to isolate special dwellings. LFS is not interested in these collective sub-blocks, so the inset is usually deleted unless it forms a PSU boundary as well. Often, the number of insets can be reduced by simply increasing the map size. Once map size and insets are settled, the map is recreated using the LFS Mapper software. Detailed clerical verification steps are outlined in Lindeyer (2004).

In some situations, the block numbers and starting points are eliminated from the map. In rural areas, the number of blocks is sometimes excessive, making it difficult for the automated sequencing program to find a suitable sequence of blocks. In rare cases, the geography of the PSU corrupts the placement or sequence of block numbers. Only in AR Group 3 can we drop block numbers and starting points, since the other two groups require a visible sequence to relate to the listing from the Address Register.

Street names are labelled on the diagram using Arc/Info default procedures. Sometimes, a critical street is missing a name. Occasionally, alternate sources are used to add the name manually.

Civic numbers are not generally used on the map. It has been an LFS tradition to list without the aid of civic numbers, in part to encourage a more thorough investigation of the area, and in part due to the poor quality of civic numbers in some areas. Civic numbers are manually added where required to identify an imaginary boundary crossing a street. Imaginary boundaries imply that no physical feature is available to identify the limit of the PSU. A civic number on the street at this point is useful for determining the PSU boundary. In many cases however, the civic number is not available. The interviewer has to guess based on the distance travelled or knowledge of the municipal boundary for example.

The Address Register Data Quality check identifies odd situations that may require additional notes on the map. In some large apartment buildings, the description of the contents is added to the diagram as inclusions or exclusions, as the case requires. Large PSU splits also require annotation, as described below.

In exceptional cases, the oddly-shaped PSUs generated by the reconciliation to the latest road network have to be manually altered to provide an easily-identifiable boundary with a count of households similar to the original PSU design on either side of the problem.

3.2.5 Large PSU split

The formation of PSUs attempted to produce areas containing 200 households. Typically the GArDs program left one or more PSUs with a large count in each unit of work. Working with Census blocks as the building unit for PSU formation, some blocks are large at the outset. At times the count exceeds 450 households and is deemed too large for one PSU. In addition, Address Register updates produce a number of PSUs with a large amount of growth, whereby the household count is updated. Some of these also exceeded the 450-household limit. For this reason the unit delineated by GArDs is called the LFS_Geocode. The units are only labelled as PSU after splitting. The larger

LFS_Geocodes are split into two or more PSUs prior to sample selection. Most LFS_Geocodes have only one PSU. Only 1.5% of them require splitting. Three forms of splitting are generated as follows. The first two forms of splitting provide a new household count for each component PSU.

The simplest split is where one civic address, typically an apartment building, conveniently contains about one-half of the dwellings. One PSU consists of the apartment address only; the other PSU consists of the remainder of the LFS_Geocode, excluding the apartment address. Note that the automated LFS Mapper could only map the original LFS_Geocode. The *F01 Cluster Diagram* is manually annotated to clarify which PSU is to be listed. Rarely, more than one apartment address is used to split the LFS_Geocode into three PSUs. A total of 436 LFS_Geocodes are split by civic address prior to sample selection, 24 of them into three PSUs.

The second-simplest split uses a street in its entirety and is called a blockface split. One PSU consists of all dwellings addressed to this street, and the other consists of all dwellings not addressed to this street. In some PSUs, the street could have multiple intersections, making for an odd boundary to be drawn on the map. Manual annotation clearly indicated all addresses on the street. Some 260 LFS_Geocodes were split by blockface, all into two PSUs.

The final mode of splitting cannot be generated in an automated fashion as is the case for the first two. The total household count is evenly divided by the number of splits, without knowing whether or not it is possible to generate such a split. Upon selecting the first of these splits, a manual division of blocks or blockfaces is performed and retained for future introductions of the other PSUs in the same LFS_Geocode. An even division of households is the primary goal in these "virtual" splits, although a variance of up to 20% is allowed in the clerical procedures. Only 111 LFS_Geocodes require virtual splits, but 12 of them require from 5 to 14 PSUs each.

Manual annotation of split descriptions on the PSU maps involved is an important quality measure to ensure the appropriate area is listed

3.2.6 Send to Regional Offices

A tight schedule of shipments is required to send all the *F01 Cluster Diagram* in time for listing – see next section for a description of listing and listing maintenance. All AR groups are shipped together despite the fact that AR Group 1 PSUs are not required until dwelling selection occurs.

For the large influx of maps related to the start of the 2004 Design, shipment to the Regional Offices occurred between June and November 2004. Listing commenced in July 2004. Over 6,500 PSUs were mapped and over 30,000 individual maps were produced including insets and copies. Two copies were shipped, one for the Regional Program Manager, and one for the interviewer. With the smaller volume of PSU replacement in-between redesigns, a third copy was generated for the senior interviewer.

3.2.7 Regional Office feedback

Since the introduction of AR listing into field operations required substantial changes to the interviewer's job, a large amount of feedback was requested and received. Despite preliminary tests of AR generated listing, a number of recommendations were received and implemented.

The assignment of PSUs to AR groups was refined in light of the difficulties some interviewers had with following the sequence of blocks as prescribed on the map. In AR Group 2 PSUs, the interviewer is permitted to delete all AR listed dwellings if the sequencing is thoroughly confusing. The block sequence on the map can also be ignored in this situation.

The Prairie Region noted the lack of the Township, Range and Meridian designation on their maps. Several managers requested colour maps and/or topographic contours to assist in finding landmarks. Colour has been added to the LFS Mapper software requirements.

The quality of road networks became an issue in many areas. Ancient roads or railways that have not existed for many years remain in the geography database. Other roads are missing even though they are not very new.

GArDS worked with a snapshot of the road network from the 2001 Census. This road network was updated early in 2004 requiring a reconciliation effort whenever a shifting street also affected a PSU boundary. A dozen or so PSUs have odd shapes and/or are discontiguous due to imperfections in this reconciliation.

Future updates to the road network will also require reconciliation that may change active maps.

3.2.8 Address Register and road network updates

Plans are in place to improve the listing on a regular basis by adhering to the latest Address Register. The AR

has moved on to service the 2006 Census, and hence uses 2006 Census geography, including the blocks required for delineating PSUs. A new reconciliation is required to update current PSU boundaries to the new geography. About 400 PSUs have considerable discrepancies between old and new geographies, if the boundary has to snap to the new blocks. A handful of active PSUs may require listing updates to deal with changes in the area to be listed.

Only once reconciliation is complete can the steps to using the latest AR be attempted. Sequencing of the blocks within PSU, blockfaces within the blocks, and AR dwellings within the blockfaces must be redone. A new extraction follows with subsequent analysis for AR Group identification. Map size is unlikely to change, but new inset coordinates may be required due to the new road network in some areas. Every update to the AR may require these steps.

3.3 Listing and listing maintenance

3.3.1 Introduction

Listing is the process that identifies and captures all dwellings within a PSU boundary, as displayed on the *F01 Cluster Diagram*. Proper translation of the map contents in relation to physical features on the ground is paramount. Dwelling addresses or descriptions are captured by the field interviewer using the LFS Listing Application installed on their laptop computers. The LFS-100 CAPI Interviewer's Manual (2006) has instructions on both listing and working with the listing application. This manual required significant updates for the differential treatments imposed by the use of the AR.

The dwelling list is open-ended. During the life of the PSU, the interviewer can add new dwellings on a regular basis as the population grows. The fact that a list is open-ended not only allows for additional dwelling selection in the field, but may also require special treatment if the growth becomes a burden to the interviewer.

Listing proceeds differently according to the Address Register Group as outlined in Section 3.1. This section outlines the treatment by AR Group, the differences between Initial Listing and Listing Maintenance, as well as the treatment of collective dwellings, the options available in areas of dwelling growth, and the programs in place to monitor listing quality.

3.3.2 Treatment by AR group

AR Group 3 PSUs have no AR listed dwellings. The interviewer starts with a blank listing and must identify all dwellings within the PSU.

AR Group 2 PSUs have most dwellings added from an AR derived list. In a list verification procedure these dwellings must be confirmed as belonging to the PSU. Extraneous dwellings are deleted. Dwellings in the field that are missing from the list are added. Inaccurate addresses are modified, especially where the unit number is fictitious or unclear.

AR Group 1 PSUs have almost all dwellings added from an AR derived list. Unlike PSUs in Group 2, those in Group 1 are not subjected to listing verification prior to dwelling selection, *i.e.*, the AR list is used for initial dwelling selection without the benefit of field updates. Once interviewing starts, listing updates are expected to correct any discrepancies between the AR list and the actual dwellings within the PSU boundary.

The AR Group determines the initial method of listing only. Once a list has been field-verified, the original AR Group becomes irrelevant. At the moment, there is no feedback to the Address Register for any changes made to an AR-derived list. Various studies are possible to address the differences encountered.

3.3.3 Initial listing

Initial listing is the process whereby a dwelling list is prepared for the first sample selection in a PSU. The creation of an AR Group 1 list that is sent directly for sample selection is part of the Initial Listing process, although no field verification takes place.

PSUs assigned to Group 3 undergo Initial Listing in the field. PSUs assigned to AR Group 2 are sent to the field for Initial Listing verification. No sample selection can take place until the field interviewer returns the updated list for addition to the Head Office database.

The LFS Listing Application treats Initial Listing and Listing Maintenance very differently. During the initial listing, the field interviewers can very easily add, delete and move dwellings within the list. Address lines are renumbered with every change to the list. At this stage, no dwellings have selection data, hence renumbering does not affect the historical record of sample selection.

More details are available in a self-study guide (LFS Initial Listing for AR Clusters 2004) that was created to assist interviewers in the transition to the new method of listing PSUs assigned to AR Group 2. The contents of the self-study guide were transferred to the LFS-100 CAPI Interviewer's Manual (2006). See the manual for complete details on listing.

3.3.4 Treatment of collectives

The listing of collectives is not as clear-cut as with privately-occupied dwellings. Normally Head Office staff determine the appropriate number of dwellings to list for a collective. Due to the limited staff available and the large volume of PSUs shipped for listing each month during the redesign, the field interviewers were given the responsibility of determining the number of dwellings to list for each collective themselves, based on a set of predefined criteria (see below). Recommendations by type of collective should cover 99% of cases.

There are two main criteria for listing collectives. First, inmates of institutions are not part of the population covered by the LFS. An initial determination must be made as to whether or not the institution contains residents who are there involuntarily (*e.g.*, jails) or because of infirmity (*e.g.*, nursing homes). Generally only the owner's residence and any staff residences if applicable would be listed. New types of collectives related to seniors have a complex mix of simple apartments, rooms with meals provided, full-time nursing care and palliative care. The more complex examples require Head Office assistance to determine the correct listing procedure.

The second main criterion is the likelihood of reaching one or more respondents with no usual place of residence elsewhere. Motels with long term tenants, boarding houses, staff residences in various campgrounds, inns and hospitals are included in the survey. Where several tenants are rooming and sharing common areas, three or more rooms may be assigned to one listing line.

3.3.5 Listing maintenance

Once a PSU has been sampled to the dwelling level, the regular updates to the address list are called listing maintenance. Before the first sample of dwellings, an initial list will have been created and possibly verified in the field. After sampling, the rules for listing updates become more stringent to preserve the historical record of sample selection. A dwelling cannot be deleted, only deactivated, with some reason for the deactivation indicated in the dwelling description. Dwellings can be moved but the listing line is not updated. Instead, a separate print sequence is maintained to record the interviewer's preferred listing order. To ensure the interviewer takes extra care when maintaining a list, the listing application requires additional key strokes to initiate and confirm listing updates.

Since the PSU list of dwellings is open-ended, additions are made as new dwellings are constructed. A portion of the new dwellings are selected by the listing application during Interview week, using the PSU sampling rate. These Interviewer Selected Dwellings (ISD) generate cases for the CAPI interviewer to complete.

There are two forms of ISDs created during listing maintenance. Firstly, new structures added to the end of the list are sampled using the inverse sampling rate 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 inverse sampling rate.

The second form of ISDs are known as multiples. During the process of interviewing within a selected dwelling, the interviewer may determine that separate dwellings exist within the structure, typically basement or upper units not readily evident from the street. If the dwelling list does not contain the extra units as separate lines, then those dwellings may have missed an opportunity to be selected over the lifetime of the PSU. To compensate for the missed opportunities in this and any other similar unresolved cases, all missed units are selected along with the original dwelling. They are added to the list as multiples of the originally selected dwelling and the application generates cases for each multiple.

Listing maintenance is usually performed once or twice per year in every PSU, depending on the potential for growth and the availability of the interviewer. To reduce costs, maintenance is normally conducted during the birth month of the PSU, the month in which dwellings rotate and at least some of the newly selected dwellings have to be contacted in person.

First time maintenance in AR group 1

Special consideration is required in the listing maintenance of AR group 1. The first time such PSUs are maintained also represents the first time that any interviewer has access to the AR-derived list for update purposes.

Normally, early in the first week of interviewing, most effort is directed towards completing interviews. At the very least, selected dwellings in PSUs assigned to AR Group 1 must also be verified in listing. If not, errors in listing that affect the sample selection may be missed. Special attention is required for multi-unit addresses. The entire structure must be verified, to determine if one or more non-existent units were added by the AR.

More details are available in a self-study guide (LFS AR Group 1 Clusters Interviewer Manual 2004) that was prepared for all field interviewers to assist them in the special requirements for listing maintenance in the new design, especially for the treatment of PSUs in AR Group 1. The CAPI Interviewer's Manual LFS-100 (2006) was updated to include these new features.

3.3.6 Treatment of growth areas

Since PSU dwelling lists are open-ended, there is potential for extreme growth. Limits to the ability of the interviewer to maintain large lists include: a physical limit in the listing application to 999 listings; a cost associated with maintaining large lists; and a time restriction in conducting interviews for the large influx of new sample that such a large list implies. Methods to reduce the number of interviews to be conducted are called PSU sub-sampling.

Dummy PSUs

If the only problem with the growth of a list of dwellings in a PSU is that the physical limit of 999 has been reached, a dummy PSU is generated to incorporate the remaining dwellings. Typically the cluster value within the sample identification is changed from the existing value to 999. A secondary dummy, if necessary, will use 998.

Typically, the PSU involved is in an area with a large percentage of vacant dwellings. All dwellings are listed and the number of interviews is quite normal for a PSU. No PSU sub-sampling is required, and yet maintaining the large list could still be a problem because of the large number of vacant dwellings.

Modifying the application to allow more than 999 listings would require a further changes in the sample identification of LFS PSUs and subsequent changes in every system using that sample identification.

Strictly speaking, 999 is not the true limit. In order to preserve an equal number of dwellings for all possible starting points in the dwelling selection, the limit of the listing is confined to be a multiple of the dwelling inverse sampling rate (Ic).

$$\text{Last list} = \text{int}(999/Ic)*Ic$$

This limit provides a clean break between the sample selected in the original PSU list and its dummy. The dummy PSU acquires the normal rules about continuing growth and interviewer selected dwellings.

PSU Sub-sampling

Based on feedback from the field, the PSUs with large growth may impact on the ability of the interviewer to complete a birth assignment, since the number of dwellings in such an assignment is inflated. If such is the case, the PSU is sub-sampled to reduce the burden. There are two forms of sub-sampling, a simple modification to the sampling rate, and the formation of a second stage sample selection.

Simple modification of the sampling rate is used for moderate growth. The potential of additional construction is limited, and the percentage increase over the expected household count is from 100% to 200%. If growth is less than 100%, it is insufficient to warrant sub-sampling. If growth is more than 200%, consideration should be given to the formation of second-stage units and a sub-sampling of units to add an extra stage of sample selection.

Most often, it is sufficient to decrease the sampling rate by a factor of two in order to reduce the interviewer's workload. The PSU sub-weight is modified to account for the sub-sampling rate, as explained in Section 6.2.2.

Subclustering

The formation of subclusters as second-stage units requires substantial growth in the PSU of at least 200% over its expected size, a usable street network with which to form at least four SSUs, and a significant increase in the interviewer's workload.

The interviewer must provide a detailed breakdown of the new dwelling counts by blockface, including all the new streets built in this growing area. Head Office staff uses the new streets and dwelling counts to delineate four or more subclusters, attempting once again to generate sample units with approximately 200 households.

Two of the SSUs are selected, creating two new units to be listed in the field. The overall rate of sub-sampling for the PSU is typically between two and three. More than three increases the possibility of outliers significantly affecting the estimates of some variables, while less than two does not provide a reasonable reduction in the interviewer's workload. The PSU weight is adjusted for both subclusters in order to compensate for the sub-sampling. See Section 6.2.2 for additional details.

Sub-sampled units are mapped outside of the regular mapping operation and sent for listing. Dwelling selection can only commence once listing of the sub-sampled units is complete. Typically the subcluster

listing has to be done quickly since the initial listing for the original PSU is far along in the schedule of activities when the growth situation is identified.

Stratum update

Occasionally the growth in a PSU is extreme. Even PSU sub-sampling is insufficient to reduce the interviewer's workload and there may be an impact on the estimates of CV. At this point it is better to redesign the stratum. Typically, other PSUs in the stratum will also experience significant growth. Not all such PSUs are selected and the stratum may no longer contain a representative sample of the stratum population, hence the need to update the stratum.

A complete count of dwellings for all PSUs in the stratum is required. Either a more recent Address Register, a new Census or, if lacking, a field count is required. In addition, where the growth of a PSU is significant, a detailed breakdown of the dwelling count by blockface is required in order to form secondary sampling units (SSUs) as explained above.

The stratum update program is then implemented with the new dwelling counts, including the newly formed SSUs. 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. Newly selected PSUs require listing, but typically 60-70% of clusters already listed are retained in the new sample. 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 by the latest count of dwellings that is not reflected in the active sample.

Early in the current design, a case for stratum update was required in Calgary.

3.3.7 PSU yield monitor

A number of PSUs encounter difficulties during mapping and listing. The interpretation of the map may lead to erroneous lists. The PSU Yield Monitor identifies PSUs that have too few or too many households. Too few households suggests an area with demolitions, incorrect boundaries specified by the PSU formation, or difficulties detecting these boundaries in the field. Too many households usually indicates areas of growth, but may also indicate cases of large PSU splits that were not delineated correctly on the map, or not interpreted correctly in the field. PSU splits are explained in Section 2.4 and Section 3.2.

Field follow-up attempts to justify or correct discrepancies. Explanatory maps or revisions are issued if the original boundaries are unclear. Additional details available with the latest road network, including new civic ranges, may also assist in delineating the correct set of dwellings to list.

On occasion, corrections are required in the AR extraction, not so much to align the current PSU selection with the desired set of dwellings, but to provide comments concerning neighbouring PSUs that are equally affected by the changes.

3.4 Select dwelling process

Dwelling selection is the process that follows the selection of Primary Sampling Units (PSUs) in each stratum and includes: the assignment of rotation numbers to PSUs; the creation of a system of PSU rotations, not only to rotate dwellings within PSUs but also to replace the PSUs themselves; the systematic sampling of dwellings; and the stabilization of the total sample size. First, we will describe the rotation number.

Rotation number

Each month, a portion of the LFS sample is replaced. The replacement of sampling units, called rotation, occurs at each stage of the multi-stage sample design. The ultimate unit of selection, the dwelling, is replaced every six months, whereas higher-level units remain in the sample for longer periods of time. The determination of six months as the period for rotation of dwellings is a trade-off between the cost of rotation and the increase in nonresponse that might occur if respondents are asked to remain in the survey for a longer period of time.

To ensure uniform interviewer workloads and to minimize the effect of any bias due to the number of months a dwelling has been in the survey, a rotation scheme was adopted whereby one sixth of the dwellings rotate each month. This scheme is implemented by associating a rotation number between one and six with each PSU. This number determines the month in which the rotation of dwellings (their birth month) takes place. If the rotation number is 1, then dwellings in the PSU rotate in January. Since the dwellings are active for six months, rotation 1 also rotates in July to replace the dwellings selected in January. At this point the January dwellings are considered "rotated-out". Similarly for rotation 2, the dwellings rotate in February and August for rotation 3 in March and September, 4 in April and October, 5 in May and November, and finally 6 in June and December.

Off-rotation: In some situations a PSU is assigned a rotation but then introduced to the survey in a different month than the normal date of introduction for that rotation. This is called an off-rotation introduction. The date of introduction is later than normal but the expiry date remains the same, and so, fewer than six interviews are conducted in the first sample of the PSU. Off-rotation introductions are used whenever a new sample must be introduced more quickly than implied by the normal date of introduction.

Assigning rotation numbers

Rotation numbers are assigned so that the expected sample is evenly distributed across all six rotation panels. The expected sample is the yield from all sampled PSUs based on the design count of households used in creating the LFS frame. Even distribution is required at the stratum, Employment Insurance Economic Region (EIER) and province levels. Adherence to this objective implies the following:

- The workload per stratum is stable over time, as roughly equal numbers of dwellings are rotated each month.
- The sample comprises equal numbers of households having been in the sample for 1 *vs*. 2 ... *vs*. 6 occasions, mitigating time-in-sample effects in the estimates.
- The sample is effectively divided into six equally representative parts, which may be used when sub-samples from the LFS frame are desired.

To start, PSUs in a stratum are assigned rotation numbers to balance the total expected dwelling sample within the stratum. In most areas every stratum has six or twelve selections so that each rotation group can have the same number of selections. In a few strata there may be anywhere from one to five PSUs. These strata are combined where possible to create a more or less even distribution. The remote strata often have fewer than six selections and are left out of the general picture - their actual yields are very uncertain in any case.

The assignment of rotation numbers is accomplished independently within each EIER. Within an EIER, the strata are processed as they are created, each stratum relying on the cumulative expected sample, by rotation, of the strata preceding it. The assignment begins with an array of expected samples sorted by rotation. For each stratum in the list the expected sample by PSU is sorted from minimum to maximum. A cumulative count of expected samples by rotation, as collected in the preceding strata in this EIER, is then similarly sorted from maximum to minimum.

The rotation numbers are assigned by matching the PSU with the minimum expected sample in the stratum to the rotation with the maximum value in the cumulative count. At the same time, the PSU with the second-smallest value is matched to the rotation having the second-largest value in the cumulative count, and so on for all six rotations. If there are 12 selections in the stratum they are processed in two sequential batches of 6. After processing each EIER, the expected sample should not vary by more than the variation within any one stratum. A random order for these rotation numbers is assigned at the start of each EIER in order to create a distribution that is as even as possible at provincial and national levels. Note that these are expected samples based on design counts. Actual sampling will vary considerably from this expected value in some cases.

The above method applies to the random group method, described in Section 2.7. In this method each group is assigned a rotation and each PSU selection generates an expected sample size that differs from group to group. In some strata using randomized probability proportional to size, the expected sample size is the same for all selected PSUs. A random order of the six possible rotation numbers is assigned to the systematic sample of PSUs.

In strata with three-stage designs, the PSU selection stage is assigned to one rotation. Subsequent stages all have the same rotation number.

PSU rotation

PSU selection methods are described in Section 2.7. Each month, a small subset of selected PSUs reach their pre-determined life in the LFS sample and must be replaced. Slightly different methods of PSU rotation are used according to the type of design in individual strata.

Areas using the random group method: In the random group method, each group consists of a random subset of the PSUs within a stratum. The PSUs in the subset are randomly ordered and one PSU is selected with probability proportional to the PSU size. The PSU size is related to the number of households as determined by the previous Census. The random number used to select the *initial* PSU also determines the number of systematic samples to be drawn from that PSU, known as its life. Once the requisite number of samples has been interviewed, the PSU rotates.

The random retention periods for initially selected PSUs are necessary to ensure that initial probabilities of

selection of PSUs are preserved over time. If, for example, initially selected units are retained until exhausted (that is, until all systematic samples of dwellings are used), this would eventually result in a sample with an overrepresentation of larger PSUs. In addition, since many PSUs within a stratum have similar sizes, a large number of PSUs will be exhausted at the same time, creating a burden for field listing operations.

PSU rotation is carried out by proceeding to the next PSU on the randomized list of PSUs in the group. If the PSU rotation proceeds to the end of the list, the selection reverts to the first PSU on the list. Eventually, the PSU rotation returns to the initially selected PSU, but in almost all strata, the LFS design will be replaced before that happens.

The first replacement PSU for an initially selected PSU remains in the sample until all dwelling selections within the PSU are exhausted, subject to the minimum life rule explained in Section 2.7. Lengthening the life of the initial PSU selection is negated by shortening the life of the first replacement PSU selection. As a result, the second unit (*i.e.*, the replacement PSU) stays in the sample until its regularly scheduled rotate-out month, as if the minimum life adjustment for the initial PSU had not taken place. Subsequent PSU replacements remain active in the sample until all possible dwelling samples have been sampled.

For every PSU, selected or not, another number between one and the PSU inverse sampling ratio (ISR) is generated at random. This number determines the random start for the systematic selection of dwellings within the PSU. The systematic sample of dwellings generally has one more dwelling in the first start than in the last. Randomizing the initial starting point avoids the resulting gradual reduction in sample size as clusters age. The start value advances with each sampling occasion within the PSU, returning to 001 only after passing the maximum value equal to the ISR.

Areas using randomized PPS systematic sampling: Rotation of sample units proceeds more or less as described in the previous paragraphs. Instead of selecting one PSU per random group, there can be one or several PSUs selected from a randomly ordered list of the entire stratum. With n as the number of PSUs to select and I_s as the stratum ISR, a random number r is selected from 1 to $n * I_s$. The PSUs are initially selected using a sequence of starting points $r + k * I_s$ for $k = 0$ to $n - 1$. These starting points determine the retention period of each *initially* selected PSU. As above, a separate random number from 1 to the PSU ISR is chosen to start dwelling selection.

An example will illustrate this method. Here is a randomly ordered list of five PSUs. The stratum ISR $I_s = 84$ with $n = 2$.

Table 3.3 LFS randomized PPS systematic

PSU	HHD	ISR_u	ISR	CM	S	Life
3	153	24.637	25	25	…	…
4	241	38.808	39	64	61	4
1	224	36.071	36	100	…	…
5	218	35.105	35	135	…	…
2	207	33.333	33	168	145	24

Note: CM is cumulative, S is random start. See Appendix A.2 for other abbreviations.

The household counts (HHD) from the Census are totalled to yield 1,043 and divided by $I_s * n = 168$ to produce an expected sample size of 6.208 per selected PSU in this example. This expected sample size is divided into each PSU HHD count to yield the un-rounded ISR_u. The ISR_u are then rounded and adjusted to yield the PSU ISR column. Adjustments from straight rounding are required so that the sum of ISR remains equal to $I_s * n$. The ISR column is the true size measure used in the PPS sampling procedure. The CM column cumulates ISR from top to bottom. The PSU selection is derived from a random start r between 1 and 168, say $r = 61$, in our example. Examining the CM column, choose row i where $CM_{i-1} < 61$ and $61 <= CM_i$. Row 2 satisfies these constraints. For the second selection, $n = 2$ whereby $r + k * I_s = 61 + 1 * 84 = 145$ and row 5 is selected as shown in column S. One can see that a necessary condition for the selection of more than one PSU is that each ISR must be less than the stratum ISR. This condition applies easily if the HHD counts are more or less equal across PSU in the stratum. The life of the PSU selection equals $ISR_i - (S_i - CM_{i-1}) + 1$. The first PSU remains in-sample for 39 – (61 – 25) + 1 = 4 occasions or starts. The second PSU remains in-sample for 24 starts.

Another way to think about this scheme is that the randomly ordered list of PSUs is a partially-random ordered list of all of the possible samples within those PSUs, each sample corresponding to a starting point for dwelling selection. It is only partially random since all of the possible starting points from the same PSU are together. We then select a systematic sample of starting points from this list.

Areas using 3-stage sampling. There are a few urban strata that use 3-stage sampling. The same sample-replacement scheme applies to each stage of sampling. Dwelling selection in the Second Stage Unit (SSU)

continues until the retention period of the SSU has elapsed and it rotates into the next SSU. SSUs rotate within the PSU until the retention period of the PSU has elapsed and the PSU rotates. The new PSU selects SSUs within itself to start a new sequence. Rotation numbers are assigned at the PSU level in order to minimize costs. These strata have features similar to remote areas. Since remote areas may only be visited once every six months, it is more efficient if the entire PSU has the same birth month.

Dwelling selection

Dwelling selection occurs on a monthly basis, one rotation at a time according to the birth month of each rotation. The systematic sample of dwellings requires the PSU ISR and a starting point. (In a three-stage design, the SSU ISR is the relevant one for dwelling selection.)

Dwellings are selected from the PSU using a list of addresses generated by the Address Register and/or field listing of the PSU. Replacement PSUs must be completed and added to the listing database prior to the first dwelling selection, otherwise it is called a late-listing. The systematic sample is generated by selecting listings starting with the line number equal to the start r and selecting additional lines thereafter by adding the PSU ISR I. We select $l = r + k * I, k = 0, 1, 2, \ldots$ until the line number exceeds the number of lines available on the listing.

The first selection of dwellings in a PSU uses the random start chosen at the PSU selection stage. For each subsequent selection the random start for the PSU is incremented by one, until the life of the PSU has reached its maximum. If the incremented value exceeds the PSU ISR, the start reverts to 1.

For initially selected PSUs it is not necessary that every dwelling be selected prior to replacement of the PSU.

Sample size stabilization

PSU listing is open-ended in the field. Additions to the list generate new sample according to the sampling rate of the PSU. The stabilization program was instituted in the 1974 design to limit the growth of the LFS sample. In the current design, the 3% sample size reduction required to fund the redesign was implemented in part by using the stabilization program.

The number of newly sampled dwellings (births) selected each month is limited to a previously set value called the base figure. Any dwellings exceeding this limit are randomly dropped. The remaining dwellings are assigned a stabilization weight to compensate for the drop.

The base figures and sample drops are computed by area and rotation. The monthly LFS sample has only one on-rotation number according to the birth month. The monthly calculation of stabilization weights implies a separate weight for each rotation number.

Each area starts with an Employment Insurance Economic Region (EIER). Within each EIER separate areas are created for the special strata that may be present, including Remote, Aboriginal, High-income and High-vacancy. Some EIER are equivalent to Census Metropolitan Areas (CMAs) in physical area, hence stabilization areas respect CMA boundaries.

Not all areas are included in stabilization. Remote and high-vacant strata, the northern territorial sample and a few aboriginal areas with very small expected sample are not included in the stabilization program. Their actual monthly sample is too variable to warrant stabilization.

The drop of dwellings is performed according to the following algorithm. Off-rotation PSUs and PSUs with large growth are identified first.

Off-rotation PSUs are those introduced in a month other than that dictated by the rotation number. They are completely excluded from stabilization. These PSUs properly belong to a different birth rotation month, a different sampling month.

Large growth PSUs are sub-sampled as explained in Section 3.3. These PSUs are assigned a weight to compensate for the sub-sampling. Sub-sampled PSUs are included in the count of dwellings for the stabilization area but they are ineligible for dropping as explained below. The stabilization weights are not applied to sub-sampled PSUs to avoid any increase in the PSU weight already present. PSU-level weights are further described in Chapter 6.2.2.

PSUs that are not stabilized have a default stabilization weight of 1. The stabilization weight applied to the remaining PSUs is calculated within each area (and rotation) using

$$w = (N - C)/(B - C)$$

where:

- w is the stabilization weight,
- N is the original count of selected dwellings,
- B is the base figure,

– C is the count of selected dwellings in sub-sampled PSUs. Typically C is zero.

The number of dwellings to drop is N-B. The drop is systematically applied to the N-C dwellings using a random starting point within each stabilization area.

If N is less than B then no stabilization is applied to that area for that month. In other words the stabilization weight defaults to 1. We cannot increase the sample size when the base figure B is larger than the actual sample. This happens when new PSUs are not listed in time for sampling, or the population of an area decreases over time. In addition, the variability by rotation increases over time and some rotations will have smaller than expected samples.

Dwellings selected in the field due to growth in the cluster have no chance to be included in the stabilization program. These dwellings are excluded from the stabilization weight. Dwellings selected due to misidentification of multi-unit structures as single residences can lead to all units being selected upon sampling the main residence. Such "multiples" can be given the stabilization weight, in effect appropriating the weight of the main residence.

Other surveys also use the stabilization program. For example, the Survey of Household Spending selects an independent set of households from the active set of PSUs in the LFS. Complete rotations are normally reserved for their use. Stabilization reduces the sample yield to a fixed total.

Development of base figures

The LFS design is based on an allocation of households distributed across more than 1,000 strata. The dwelling selection process is based on listed dwellings in all PSUs selected in these strata. In order to stabilize these dwelling selections to the required set of household allocations, the base figures are calculated as an estimate of the number of dwellings that will ultimately yield the required household count. Base figures are calculated on a monthly basis. The factors that impact on this calculation include the sample allocation in households, the rotation imbalance, if any, the conversion to dwellings and the presence of non-stabilized areas.

Sample Allocation: The starting point in the calculation of base figures is the allocated sample from the design of the LFS frame. For each stabilization area, this is the number of occupied households (respondent and non-respondent) expected in the survey.

Rotation Imbalance: Due to the process of assigning rotation numbers as explained above, the number of expected households may not be exactly equal across the six rotations. Differences rarely amount to more than 1%. Using an expected sample by rotation implies that stabilization weights are more uniform across rotations. As the design ages, the relative merit of this imbalance becomes less important and a simple division by 6 is used.

Conversion to Dwellings: The final stage of sampling of the LFS is based on a list of dwellings for each selected PSU, six weeks prior to identifying whether or not the dwelling is a household. The desired household sample size must be converted into a base figure of dwellings for each stabilization area. The major factors in this conversion include the vacancy rate, new dwellings selected in the field and dwellings that are deactivated during listing maintenance.

The major difference between the number of households and the number of dwellings is the set of responses included in the categories "Vacant", "Seasonal", "Under Construction", "Business" and "Residents not eligible".

Such out-of-scope dwellings comprise 5% to 20% of the total number of dwellings, depending on the stabilization area. For each stabilization area, an estimate of the vacancy rate is calculated based on the latest survey results available. At the start of the design, a more approximate estimate is obtained from Census records with a small adjustment for the known differences between Census dwellings and LFS dwellings. A new design may take a year before proper base figures can be calculated accurately.

Sometimes a new PSU is not listed in time for sampling. The list is completed during the survey week and generates a sample in the field known as Interviewer Selected Dwellings (ISDs). In other PSUs, additional dwellings may be found by the interviewer during the survey week that also generate ISDs. These late additions are not available to the stabilization program. Hence, a reduction is made to the base figure in order to allow for these extra households after sampling. The additional dwellings average about 1%-2% of the initial sample size. Estimates of ISDs are based on the most recent survey results, but are limited to 4% in any one stabilization area to avoid outliers.

Listing errors, demolished buildings and relocated trailers can reduce the birth sample of dwellings by about 1% during the survey week. Base figures are increased to compensate for this loss. Estimates of

deactivated dwellings are based on the most recent survey results, but are limited to 2% in any one stabilization area to avoid outliers.

The above three factors suffer from the fact that past performance is not always a good indicator of future values. The final monthly sample size in households will fluctuate slightly when compared to the sample allocation required. The largest factor is the vacancy rate, which may have a seasonal component, and may change unpredictably with each new replacement PSU.

Non-stabilized Areas: The sample size in non-stabilized areas amounts to about 1.3% of the total sample. Sample sizes in non-stabilized areas are allowed to grow while stabilized areas are maintained relatively constant. Fortunately, non-stabilized areas are mostly rural and remote areas with little prospect for growth. There is no counterbalance in stabilized areas for the potential growth in non-stabilized areas. Typically, problems with listing, and with travel to the more remote areas, keep the sample size smaller than expected.

Stabilization reweight

The stabilization weight, used to compensate for dwellings dropped from the sample, is recalculated after the drop is completed. Not all strata in one stabilization area have the same stratum ISR. Large differences between strata may occur which will have an impact on the applicability of this one stabilization weight for all PSUs in the area. The weight is recalculated as if the drop was performed separately for each set of strata with the same ISR within the area. These small sets of strata cannot be used as stabilization areas directly since the sample size will be too small for a viable estimate of dwelling base figures.

Chapter 4 Collection

Introduction

The data used to produce monthly employment and unemployment estimates is obtained every month by contacting the surveyed households (see Section 1.3 for more information on coverage).

The survey schedule and the collection methods used are described in the next sections.

4.1 Survey schedule

Data is collected by a team of roughly 1,500 interviewers working out of the regional offices.

The survey cycle for a given month begins as soon as the data processing from the previous month is completed. The collection activities follow a strict timetable established according to 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 15th day of that month. Interviews begin on the Sunday of the collection week and generally continue until Tuesday of the following week. On rare occasions, collection is extended by one day.

The data collected by the interviewers is transmitted to head office for processing.

4.2 Collection method

For collection purposes, there are two types of households in the LFS. One-sixth of the monthly sample consists of households in their first month of the survey. These are called "births". The remaining five-sixths is made up of households that are between their second and sixth month of the survey. These are called "subsequents". It should be noted that in the event of nonresponse or a complete change in the household occupants, collection procedures for births apply to households the first month they are surveyed.

LFS interviews are conducted using two collection methods, computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI). Historically, CAPI has been used for households in their first month of the survey, with interviewers visiting in person to conduct the interview. Interviews with subsequents are normally conducted using CATI from CATI call centres. If requested by the household members, the subsequent interviews can be done in person using CAPI.

Subsequent interviews used to be done by telephone by the CAPI interviewers. Thus the same interviewer normally conducted both the birth and subsequent interviews for a gvien household. In mid-2000 centralized CATI was introduced for subsequent interviews, meaning that the interviewers for the subsequent interviews were different from those who conducted the birth interviews, although in many cases the same interviewer conducted all of the subsequent interviews for a household. From March to September, 2005 a CATI call scheduler was phased into use. The call scheduler automatically selects cases for interviewers from a central pool as they work, so that a household may now be contacted by any interviewer.

A major change was introduced in 2004: births in urban areas for which a telephone number is available are now surveyed using Telephone First Contact (TFC). These births are contacted using CATI by interviewers in the CATI worksites. This approach was introduced to reduce the collection costs associated with an initial personal interview.

During the birth interview, the interviewer collects 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 provided in the initial interview for subsequent months.

Responses for household members are usually given by a single well-informed member of the household. This is called "proxy response" and is applied when it would be too time-consuming and costly to make several visits or calls to obtain the information directly from each household member. Approximately 65% of the data in the LFS is obtained using this method.

To maintain high LFS response rates, different types of letters are sent to the selected households. 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.

4.3 Telephone First Contact

Telephone First Contact (TFC) was introduced in November 2004. It is used to make initial contact with some births by telephone from the CATI call centres.

TFC is used for households selected in urban areas only. Administrative lists for these areas are updated on a regular basis and the addresses tend to have a standard form, which provides a better match. When most contacts are made by telephone, collection costs are considerably reduced.

Every month, around 9,000 birth households are introduced into the sample. Of these births, around 6,000 are selected in major urban areas. Births in urban areas will be matched with the administrative telephone number files. A telephone number is found for approximately 60% to 70% of these births, though this percentage may vary. In 2007, roughly 4,000 births per month were included in TFC.

Procedures are in place to ensure that at least one attempt is made to contact each TFC household during the first two days of collection, Sunday or Monday.

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 the household contacted is not the right one, other sources are searched to find a valid telephone number for the selected dwelling. If a valid number is not found, the case is transferred to a field interviewer who will go to the address in question.

If no contact is made for two consecutive months, the case will be transferred to a field interviewer for its third month in the survey. Dwellings can also be transferred to field interviewers during collection, if the respondent requests a personal interview.

Chapter 5 Processing and imputation

Introduction

There are two types of nonresponse in the LFS: person nonresponse and item nonresponse. Person nonresponse occurs if it is not possible to obtain any survey information for a person due to a refusal or non-contact. Item nonresponse occurs when it is not possible to obtain information about one or more of the questionnaire items. In the current nonresponse treatment strategy, different methods are used to handle these two types of nonresponse. Item nonresponse is dealt with through a mixture of three imputation methods: Hot-Deck (HD) imputation, carry-forward imputation and imputation by deduction. Person non-response is dealt with using either hot-deck imputation or nonresponse weight adjustment, depending on the response history of the household containing the person.

In HD imputation for the LFS, 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. An innovation in the HD imputation strategy used by the LFS is to use previous month values (perhaps imputed) of some variables in defining the imputation classes. This innovation, sometimes referred to as longitudinal hot-deck imputation, was implemented in January 2005 based on research by Bocci and Beaumont (2004).

The LFS has several design and collection features that have an impact on the current nonresponse treatment strategy. They are summarized below:

- The sample is selected according to a stratified multi-stage sampling design and is divided into six rotation groups. By design, each rotation group is representative of the entire population.
- Selected dwellings remain in the sample for six consecutive months (the term selected households should be understood to mean households currently occupying selected dwellings). Each rotation group contains households (dwellings) that have been in the sample for the same number of months. For any given month, there is a birth rotation group where the households have been in the sample for only one month, and five other rotation groups where the households have been in the sample from two to six months. As a result, the sampled households are common from one month to the next for five rotation groups out of six.
- The main variables of interest are categorical and are related to the labour force status of household members. Variables related to earnings are secondary and are only collected in the first month a person is in the sample or if the person's job situation changes. For the other months, the information related to earnings is not collected but is simply carried forward to reduce respondent burden.
- Collection is done mostly by telephone, except for the birth rotation group for which a substantial proportion of interviews are done in person.

The remainder of this chapter describes the steps in processing and imputation in more detail. Additional details of the LFS imputation methodology can be found in Lorenz (1996) and Bocci and Beaumont (2004). The former paper describes the HD imputation system specifications. The latter focuses on longitudinal hot-deck imputation.

5.1 General data processing steps

Data processing, from collection to weighting, can be divided into the following six steps:

1) Receipt of data from the regional offices and first phase of editing
2) Industry and occupation coding
3) Consistency editing
4) HD imputation
5) Derivation of variables
6) Weighting and production of a clean microdata file (the TABS file)

In the first phase of editing, imputation by deduction and some carry-forward imputation is performed. Imputation by deduction is applied when a missing value can be deduced logically from the responses to other variables. Carry-forward imputation consists of transferring historical information to the current month. At the same time, different flags such as those indicating the type of nonresponse and those indicating the need for HD imputation are set. Consistency edits are applied in step 3. Records that fail the consistency edits are inspected manually to remove the inconsistencies. Socio-demographic variables are imputed before HD imputation is performed. These variables are used to form imputation classes.

5.2 Imputation pre-processing

Before the actual imputation of missing values using HD can be performed, some pre-processing steps are followed. First, persons are 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 do not belong to group A, and who have no missing values and are internally consistent after the first phase of editing (where imputation by deduction or carry-forward imputation was performed). These persons have a complete record but are excluded from being potential donors. The remaining persons, the recipients, form group C and need HD imputation. A second pre-processing step converts all earnings data to an hourly basis. This ensures that the value imputed for earnings is consistent with the value for the number of hours collected for a recipient.

The last pre-processing step is the identification of outlier earnings. Earnings values that are either extremely high or low are deemed suspicious, hence they are set to missing and are imputed. Individuals who have very high or very low earnings, without being extreme, keep their reported value of earnings but are excluded from being potential donors. These persons are also assigned to group B. The outlier detection method used is the quartile method. Essentially, it identifies earnings that are either much larger or much smaller than the median. Records with values above a set threshold or distance from the median are considered outliers. Different thresholds are used to identify outliers that are suspected of being incorrect and those that are excluded from being potential donors. More details about outlier detection in the LFS are given in Lorenz (1996).

5.3 Imputation for item nonresponse

Once all 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. The procedure is applied in such a way that the data of a recipient after imputation satisfy consistency edit rules and validity edit rules (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 donors using simple random sampling without replacement until a donor that satisfies all the edit rules is found.

The initial imputation classes are formed by crossing the following categorical variables, ordered by importance:

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).

These variables were judged in empirical studies to be in decreasing order of importance for explaining the labour force variables. A more detailed description of the variable categories is given in Section 5.5.

Note that the variables LMSLFS3, LMSLFS7 and LMINDG refer to values from the previous month.

The use of variable TPATH is somewhat complex. To explain its role, first note that the labour force status variable LFSSTAT can take one of 7 values, which correspond to the first 7 values of variable TPATH. One of these 7 LFSSTAT values is assigned to each donor. For the recipients, the value of LFSSTAT may not be known. However, there may be enough information to exclude some of the 7 possible values. The role of variable TPATH is to ensure that only valid values are imputed to recipients. This is achieved by assigning only one value of TPATH to each recipient and by replicating each donor by its number of valid TPATH values. At the end of the imputation step these replicated donors are removed. For example, assume that a donor has LFSSTAT = 2. This donor has thus TPATH = (2, 8, 10) as valid values. It is therefore replicated three times; each replicate is assigned only one of these three TPATH values.

Imputation is performed in each class that contains a sufficient number of donors. Two constraints are used to determine if a class has a sufficient number of donors:

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

ii) Each class must contain at least 3 donors.

If one of these constraints is not satisfied, then imputation is performed again by removing the least important variable (DWELRENT) when forming the imputation classes. 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 – see Section 5.4.

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 whole record imputation is performed. 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 consistency edit rules are checked. If all edit rules are satisfied after this second attempt then the imputation process for this recipient is completed. Otherwise, whole record imputation is performed using the last attempted donor.

During the imputation process, output flags are written to the output file. These are useful for producing different tables that can be used to monitor and improve the imputation process. Detailed specifications of the above imputation strategy can be found in Lorenz (1996).

5.4 Imputation for person nonresponse

Person nonresponse occurs when a person within a household is nonrespondent or when the entire household is nonrespondent. If a household is nonrespondent but responded in the previous month, then all the labour force variables of each person belonging to the household are imputed. If the household was also nonrespondent in the previous month then a non-response weight adjustment is performed.

Prior to January 2005, a combination of cross-sectional HD imputation and carry-forward imputation was used to deal with person nonresponse. Carry-forward imputation of historical information was used for socio-demographic variables. It was also used for all other variables if the nonrespondent had responded in the previous month. If the person had not responded in the previous month, but had responded in the past, then cross-sectional HD imputation was used instead of carry-forward imputation.

The main issue with carry-forward imputation is that it underestimates month-to-month changes. On the other hand, cross-sectional HD imputation has a tendency to overestimate month-to-month changes. To overcome these problems, new longitudinal hot-deck variables are used in the definition of the HD imputation classes (see Bocci and Beaumont, 2004). This was implemented in January 2005 to deal with person nonresponse when historical information is available. Before HD imputation, all socio-demographic variables are imputed using carry-forward imputation as these variables are not expected to change significantly from one month to the next.

The variables used to form HD imputation classes for person nonresponse are given below in order of importance:

1) PROV (10 categories)
2) AGEGP1 (5 categories)
3) LMLFS7 (7 categories)
4) SEX (2 categories)
5) ABQ1 (2 categories)
6) IMM (3 categories)
7) EIER (Employment Insurance Economic Region – 55 categories)
8) EDUC (2 categories).

Note that donors and recipients for HD imputation are persons even when dealing with entire household nonresponse. Note also that validity and consistency edit rules are irrelevant as whole record imputation is performed (the donor already satisfies all rules). Finally, note that missing values of recipients are imputed using donor values of the current month even though imputation classes are based on values of the previous month.

5.5 Definition of variables used for the determination of imputation classes

Age groupings (AGEGP1 and AGEGP3 are coarser groupings of AGEGP2)

AGEGP1	AGEGP2	AGEGP3	
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
4	6	2	45 to 54
4	7	2	55 to 64
5	8	3	65+

Occupation Groupings (OCC4 is a coarser grouping of OCC10).

OCC4	OCC10	
01	01	Managerial, administrative, natural science, social science, religion, teaching, medicine and artistic
02	02	Clerical
02	03	Sales
02	04	Service
03	05	Farming, fishing, forestry and mining
03	06	Processing, machining and fabricating
03	07	Construction
03	08	Transportation
03	09	Materials handling and other crafts
04	10	Never worked before or last worked more than 1 year ago or permanently unable to work

EDUC:
0 Person does not have a high school diploma
1 Person does have a high school diploma

COW - class of worker
1 Paid employee
2 Self employed
3 Unpaid family worker

STUD:
0 Not full-time student
1 Full-time student

DWELRENT:
1 Dwelling owned
2 Dwelling rented

PROV:
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:
M Male
F Female

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

MULTJOB - Did you have more than one job or business last week?
1 Yes
2 No or no response

IMM - country of birth
1 Canada
2 United States
3 Other

Aboriginal identity – North American Indian, Métis or Inuit.

ABQ1
1 Yes
2 No

Last month's labour force status**.** LMLFS3 is a coarser grouping of LMLFS7

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

LMLFS3
1 Employed (LMLFS7=1,2)
2 Unemployed (LMLFS7=3,4,5)
3 Not in the labour force (LMLFS7=6,7)

Last month's industry group.

LMINDG
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

Chapter 6 Weighting and estimation

Introduction

Estimation is the survey process in which unknown population parameters are approximated using data from a sample, possibly in combination with auxiliary information from other sources. Estimation results are used to make inferences about these unknown parameters, that is, to draw conclusions about characteristics of the complete population using only a sample of that population. Examples of population parameters of interest include population totals, means and ratios, as well as their averages over a number of survey months. An estimate[20] $\hat{\theta}$ is thus an approximation of an unknown population parameter θ and the difference $\hat{\theta} - \theta$ between these two quantities is called the total survey error. The total survey error can be divided into two main components: the sampling error and the nonsampling errors. The sampling error is due to the fact that estimates are computed using only a sample of the whole population, while the nonsampling errors are due to other causes such as an imperfect frame, measurement errors and nonresponse. In the Labour Force Survey (LFS), the sampling error and part of the error due to household nonresponse are dealt with by attaching an estimation weight, called the final weight, to each sampled person for which we have data, be they imputed or not. To simplify the discussion in this chapter, we assume that there is no nonsampling error other than the household nonresponse error, although we briefly point out that frame imperfections can be partially dealt with using the final weights.

The basic weighting principle is to weight each person by the inverse of his or her probability of inclusion in the sample. This ensures that the estimates are unbiased, or approximately unbiased, in the sense that the expectation, over all possible samples, of the survey error is exactly, or approximately, equal to zero. To evaluate the quality of an estimate and to obtain valid inferences, measures of precision such as the estimated coefficient of variation are usually computed. The coefficient of variation is defined as $CV(\hat{\theta}) = \sqrt{V(\hat{\theta})}/\theta$, where $V(\hat{\theta})$ is the variance of the estimates over all possible samples. Since only one sample is selected in practice, the variance $V(\hat{\theta})$ is unknown. However, it can be estimated using only that sample (see Chapter 7), which allows us to obtain the desired measures of precision.

The way the sample is selected has an impact on the inclusion probability of households and thus on their final weight. The entire selection process can be divided into three main steps: i) selection of the initial stratified multistage sample; ii) sample adjustments in order to deal with growth clusters and iii) sample adjustments aimed at maintaining the sample size over time. These three steps and their impact on inclusion probabilities are discussed in Section 6.2. After sample selection, a weight can be computed which we call the design weight. It is often interpreted as the number of times that each sampled unit should be duplicated to represent the target population. For many reasons, such as refusals or the impossibility to contact some of the sampled households, the number of households for which information is collected is smaller than the number of sampled households. This reduction in the sample size leads to household nonresponse error. To compensate for household nonresponse, imputation (see Chapter 5) and nonresponse weight adjustment are used. Nonresponse weight adjustment consists of adjusting the design weight of each responding household 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. The weight obtained after nonresponse adjustment is called the subweight. How to obtain the subweight is discussed in Section 6.3. Finally, calibration is often used in surveys to obtain final weights. The basic idea of calibration is to find final weights that are as close as possible to the subweights while satisfying specific constraints. For the LFS, the constraints are chosen in order to i) ensure consistency with external estimates of population, ii) account for undercoverage to some extent and iii) improve the efficiency of the estimates. In addition, the LFS has been using a composite calibration estimator since January 2000 to improve the efficiency of the estimates, especially for estimates of change. Calibration and composite calibration are described in Section 6.4 along with the integrated method of weighting, which ensures a common final weight for every person within a household. In the next section, we define some basic concepts and introduce notation.

6.1 Basic concepts and notation

To start the discussion on weighting and estimation, let us first assume that the population parameter of interest is the population total

20. For simplicity, we blur the distinction between estimator and estimate.

$$t_y = \sum_{k \in P} y_k,$$

where P denotes the population at the current month and y is a variable of interest. For instance, y_k could be a binary variable indicating whether a given person k of the population is employed ($y_k = 1$) or not ($y_k = 0$). In that case, the population total t_y would represent the number of employed people in the population P. From this population, a sample of households is selected according to a stratified multi-stage sampling design and the LFS information is collected for every person in the selected households. In the absence of household nonresponse, the population total t_y could be estimated by

$$\hat{t}_y^D = \sum_{k \in s} w_k^D y_k,$$

where s represents the sample of all the people belonging to the selected households and w_k^D is the design weight attached to person k. Since there is household nonresponse, variable y is not observed for all the people in the sample s and the design-weighted estimate $\hat{t}_y^D$ cannot be computed. In this case, the population total t_y is instead estimated by

$$\hat{t}_y^{\mathrm{NA}} = \sum_{k \in s_r} w_k^{\mathrm{NA}} y_k,$$

where s_r is the subset of all the people from s who belong to a responding (or imputed) household and w_k^{NA} is the nonresponse-adjusted weight attached to person k, which we call the subweight of person k.

As mentioned in the introduction, calibration is used in the LFS for consistency reasons, to deal with under-coverage and to improve efficiency. Calibration leads to the estimate

$$\hat{t}_y^C = \sum_{k \in s_r} w_k^C y_k,$$

where w_k^C is the calibration weight attached to person k. Finally, to obtain more efficient estimates, composite calibration is used in the LFS. It leads to the estimate

$$\hat{t}_y^{CC} = \sum_{k \in s_r} w_k^{CC} y_k,$$

where w_k^{CC} is the composite calibration weight attached to person k which we often call the final weight of person k. The weighting steps are described in more detail in Sections 6.2, 6.3 and 6.4.

Often, interest is not in the estimation of a population total but a population rate

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

where y_1 and y_2 are two variables of interest. For instance, y_{1k} could be a binary variable indicating whether a given person k of the population is unemployed ($y_{1k} = 1$) or not ($y_{1k} = 0$) and y_{2k} could be a binary variable indicating whether person k is in the labour force ($y_{2k} = 1$) or not ($y_{2k} = 0$). In such a case, the population rate r_{y_1, y_2} represents the unemployment rate in the population. It can be estimated using the final weights w_k^{CC} by

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

Finally, it is sometimes of interest to estimate the average of a population parameter over more than one month, which can be written as

$$\theta = \sum_{t=1}^{T} \frac{\theta_t}{T},$$

where θ_t is the population parameter at month t and T is the number of months used in the definition of the above average. For instance, θ_t could be the unemployment rate or the number of employed people at month t. The parameter θ is called a T-month moving average parameter. It can be estimated by

$$\hat{\theta}^{CC} = \sum_{t=1}^{T} \frac{\hat{\theta}_t^{CC}}{T},$$

where $\hat{\theta}_t^{CC}$ is an estimate of θ_t obtained using the final weights at month t. In the LFS, three-month moving average estimates of unemployment rates for each Employment Insurance Economic Region (EIER) are produced every month of the survey using the three most recent months. Such estimates are more stable than monthly estimates but their interpretation is different since they estimate a different population parameter.

6.2 Design weight

In principle, the design weight of a person k is equal to the inverse of his or her probability, π_k^D, of being selected in the sample s; that is, $w_k^D = 1/\pi_k^D$. If there were no nonresponse or any other nonsampling error, this would ensure that the resulting design-weighted estimates are unbiased, or at least approximately unbiased. 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. There are different random mechanisms involved in the selection of households. Each mechanism has an effect on the resulting selection probability. In the remainder of this section, these mechanisms are described along with their effects on the design weight.

6.2.1 Basic weight

At the time of the survey design, strata are formed by grouping together geographic units. Details of the stratification can be found in Chapter 2. Within each stratum, a sample of households (or dwellings to be more precise) is selected using a multi-stage sampling design. In most strata, a two-stage sampling design is used and we will focus on these strata to explain how the basic weight is derived. Also, we will make no distinction between households and dwellings in the remainder of this chapter and we will use the term "households" to denote one or the other.

Each stratum *h* is divided into N_{1h} Primary Sampling Units (PSU). At the first stage of sampling, n_{1h} of them are selected with probability proportional to the quantity R^*_{hj}, where the subscript *hj* denotes PSU *j* in stratum *h*. The first-stage selection probability of PSU *j* in stratum *h* is thus

$$\pi_{1hj} = \frac{n_{1h} R^*_{hj}}{\sum_{j \in P_{1h}} R^*_{hj}},$$

where P_{1h} is the population of PSUs in stratum *h*. The quantity R^*_{hj} is equal to the ratio $R_{hj} = \tilde{N}_{hj} / \tilde{n}_{2hj}$ rounded to an integer, where $\tilde{N}_{hj}$ is the approximate number of households in PSU *j* of stratum *h* that is known from the 2001 Census and $\tilde{n}_{2hj}$ is the number of households that is initially planned to be selected in PSU *j* of stratum *h*. Note that R^*_{hj} corresponds to the number of rotations in PSU *hj*, and is also very closely related to the inverse sampling ratio (ISR) for the PSU defined in Equation (3) of Section 2.7. We assume for now that $\tilde{n}_{2hj}$ is constant within a stratum so that $\tilde{n}_{2hj} \equiv \tilde{n}_{2h}$. This assumption is removed later when we discuss weighting in the case where the Rao-Hartley-Cochran (RHC) method of selection is used. Selecting PSUs with probability proportional to R^*_{hj} is essentially equivalent to selecting PSUs with probability proportional to R_{hj} if these ratios are not too small, which is in turn equivalent to selecting PSUs with probability proportional to $\tilde{N}_{hj}$ since $\tilde{n}_{2hj}$ is constant within a stratum.

In each selected PSU *j*, a systematic sample of households is selected with the fixed sampling interval R^*_{hj}. The second-stage selection probability of household *i* in PSU *j* of stratum *h* is thus $\pi_{2hji} = 1/R^*_{hj}$. Note that the actual number of households selected in PSU *j* of stratum *h*, $n_{2hj} \approx N_{hj}\pi_{2hji} \approx N_{hj}/R_{hj}$, is likely to be different from the initially planned number, $\tilde{n}_{2hj} = \tilde{N}_{hj}/R_{hj}$, due to $\tilde{N}_{hj}$ being an approximate measure (coming from the 2001 Census) of the actual number of households in PSU *j* of stratum *h*, N_{hj}.

The overall selection probability of household *i* in stratum *h* is

$$\pi^B_{hi} = \pi_{1hj}\pi_{2hji} = \frac{n_{1h}}{\sum_{j \in P_{1h}} R^*_{hj}},$$

which is constant within each stratum; that is $\pi^B_{hi} \equiv \pi^B_h$, for all households *i* in stratum *h*. This overall selection probability is also called the basic selection probability. The initially planned sampling fraction in stratum *h* is $f_h = n_h / \tilde{N}_h$, where n_h is determined during the allocation phase of the survey design and $\tilde{N}_h = \sum_{j \in P_{1h}} \tilde{N}_{hj}$. Note that the initially planned sample per PSU is $\tilde{n}_{2h} = n_h / n_{1h}$. In the LFS, we wish to respect the initial sampling fraction that is determined after allocating the sample to strata so that $\pi^B_h = f_h$. To satisfy this requirement, R_{hj} are rounded to integers R^*_{hj} such that $\sum_{j \in P_{1h}} R^*_{hj} = n_{1h}/f_h = \tilde{N}_h/\tilde{n}_{2h} = \sum_{j \in P_{1h}} R_{hj}$. Since $\sum_{j \in P_{1h}} R^*_{hj}$ must be an integer, n_{1h}/f_h is rounded to satisfy the equality $\sum_{j \in P_{1h}} R^*_{hj} = n_{1h}/f_h$.

In the RHC method, the PSUs in each stratum are first collapsed randomly into n_{1h} groups. Then, each group *g* is viewed as a separate stratum and only one PSU is selected within each group according to the methodology described above. With the RHC method, the number of households that we initially plan to select in a given group *g* of stratum *h*, $\tilde{n}_{2hg}$, is chosen in such a way that the overall sampling fraction in each of the n_{1h} groups of stratum *h* is equal to the initial sampling fraction f_h. Therefore, $\tilde{n}_{2hg} = \tilde{N}_{hg} f_h$, where $\tilde{N}_{hg}$ is the approximate number of households in group *g* of stratum *h* that is known from the 2001 Census. As a result, it is quite likely that $\tilde{n}_{2hg}$ is not exactly the same for two different groups in stratum *h*.

Since the desired information is collected for every eligible person within a selected household, the basic selection probability of a person *k* in stratum *h* is $\pi^B_{hk} = \pi^B_h = f_h$ and his or her basic weight is

$$w^B_{hk} = 1/\pi^B_h = f_h^{-1}.$$

Such a sampling design with a constant basic weight within each stratum is called self-weighting within strata.

The basic weights would be equal to the design weights if the sampling design and the population remained unchanged. However, because the PSUs experience growth over time and the systematic sampling rate is fixed, this would lead to an ever-increasing sample size (and ever-increasing collection costs). It would also lead to large variations in interviewer workload over time and between interviewers.

To avoid this, two sampling procedures are used to control the sample size: PSU subsampling and sample stabilization. These methods change the basic selection probability of households (and people). It is thus necessary to adjust the basic weights to compensate for these sampling procedures. This is discussed in greater detail in Sections 6.2.2 and 6.2.3.

6.2.2 Cluster weight

A cluster corresponds to a PSU in strata with a two-stage design and to a penultimate unit in other strata. In urban areas, the number of dwellings in a cluster can grow substantially over time due to construction. Given the fixed sampling rate (or sampling interval) within each cluster, an interviewer's assignment size would grow substantially when this occurs. This could affect the quality of the interviewer's work in addition to his/her ability to complete the assignment. When growth in a cluster exceeds 100%, but is not too extreme, the cluster may be randomly subsampled using method I or II described below. These methods of subsampling modify the selection probabilities of households. As a result, the basic weight w_{hk}^B is modified by a cluster adjustment factor a_{hk}^P to give the cluster weight

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

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

Method I: Cluster subsampling

This method is the simplest and most common of all subsampling methods. It is used when a cluster is to be subsampled due to its growth, and neither method II nor III below applies. The cluster sampling rate is modified to reduce the number of households selected, while avoiding sampling previously selected households. The basic weights of interviewed households are multiplied by this factor. Due to outlier problems encountered by special surveys that use the LFS frame, the maximum value the cluster adjustment factor can be is 3. Also, the growth has to be sufficient to warrant a factor of at least 2 in order to use this method.

Method II: Subclustering

When growth in a cluster exceeds 200% and street patterns are well defined, the growth cluster is divided into several subclusters. A sample of the smaller subclusters is taken and then a sample of households is selected within each selected subcluster. This procedure is equivalent to adding another stage of sampling within growth clusters. It does not change the selection probability of clusters but it changes the selection probability of households within growth clusters.

Method III: Stratum updates

When growth is extreme even subclustering may be insufficient, and a stratum update is required, as described in Section 3.3.6. Updated counts of dwellings for all clusters in the stratum are required and new clusters are formed by subclustering 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 by the latest count of dwellings that is not reflected in the active sample.

6.2.3 Stabilization weight

The final stage of sampling is conducted using systematic sampling at a fixed rate. As the 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 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 EIER and the same rotation group. The set of people belonging to households that remain in the sample after stabilization was denoted by s in Section 6.1.

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

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

The stabilization adjustment factor a_{hk}^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, let us assume that we have 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.

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 method I or II in Section 6.2.2;
- 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.

6.3 Subweight

The households in the selected sample *s* are not all interviewed, due to refusals and other factors making it impossible to contact some households. Part of this household nonresponse is first treated by using a longitudinal imputation method (see Chapter 5). Then, the remaining nonrespondent households are treated by dropping them and by adjusting the design weights of responding households, including those that have been imputed, by a nonresponse adjustment factor. As pointed out in the introduction, 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 *k* belonging to a responding household in class *c* is $a_{ck}^{NA} = 1/\hat{p}_c$ and the nonresponse adjusted weight, or the subweight, is

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

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

The key to reducing nonresponse bias is to determine nonresponse classes that explain the unknown non-response 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 non-response 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 obtained by crossing the variables EIER, TYPE and ROTATION (excluding households belonging to an aboriginal or high-income class). The variable TYPE has five categories and indicates the type of stratum to which a household belongs: Remote, Rural, Urban three-stage, Urban non-Census Metro-politan Area (CMA) and Urban CMA. The variable ROTATION corresponds to one of the six rotation groups. Note that the nonresponse classes do not overlap and, together, they cover the entire population. Also, collapsing of classes is performed when a nonresponse adjustment factor is greater than two in a given class. When this occurs, this class is collapsed with another one chosen so that the nonresponse adjustment factor of the combined class is less than two. The chosen class must come from the same province, the same type of stratum and the same rotation group as the class to be collapsed, that is, collapsing is performed across EIERs. The reason for collapsing nonresponse classes is to

avoid large nonresponse adjustment factors since they tend to increase the variability of the estimates.

A new nonresponse weight adjustment methodology has been considered for potential implementation and is described in greater detail in Alavi and Beaumont (2004). In this new methodology, data collection information, such as the number of attempts to contact a household and the time of the last attempt, is used, in addition to design information, to construct classes. It was found in an empirical study that the number of attempts was the most important variable to explain nonresponse. Moreover, this information was also correlated with employment and unemployment. Therefore, it seems that this information is very useful to compensate for nonresponse. To form nonresponse classes, the score method (see Little 1986) has been proposed. This method consists of: i) modelling and estimating response probabilities using logistic regression; and ii) forming classes that are homogeneous with respect to these estimated probabilities. The advantage of the proposed score method is that the number of classes and the minimum number of responding households within each class can be easily controlled, which is not the case with the current method.

6.4 Final weight

In this section, we describe how the final weights, which are used to obtain official estimates, are derived. Composite calibration and the integrated method of weighting are the key ingredients to obtaining the final weights. This latter method is used to ensure a common final weight for every person within a household. Calibration is first discussed in Section 6.4.1 and then composite calibration in Section 6.4.2. The integrated method of weighting is described in Section 6.4.3. Finally, in Section 6.4.4, we describe how negative final weights are handled in the weighting system.

6.4.1 Calibration

Calibration is a technique that finds weights w_k^C, for all people $k \in s_r$, as close as possible to the subweights w_k^{NA}, according to some distance function, and such that calibration-weighted estimates for a vector of auxiliary variables $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}}^C = \sum_{k \in s_r} w_k^C \mathbf{x}_k$, are exactly equal to the vector of known population totals, $\mathbf{t}_{\mathbf{x}} = \sum_{k \in P} \mathbf{x}_k$. In the LFS, these known population totals, often called control totals, are in fact 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 $k \in s_r$. In the case of the LFS, this means that we must know to which age-sex group each person $k \in s_r$ belongs as well as his or her ER and CMA/CA.

More formally, calibration weights w_k^C are obtained in the LFS by minimizing the distance function

$$\sum_{k \in s_r} \frac{\left(w_k^C - w_k^{\mathrm{NA}}\right)^2}{w_k^{\mathrm{NA}}}$$

subject to the calibration constraint $\sum_{k \in s_r} w_k^C \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}$. Other distance functions could also be used (see Deville and Särndal 1992). This minimization leads to the calibration weights $w_k^C = w_k^{\mathrm{NA}} g_k^C$, where the calibration adjustment factor g_k^C is given by

$$g_k^C = 1 + \mathbf{x}_k' \left(\sum_{k \in s_r} w_k^{\mathrm{NA}} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\mathbf{t}_{\mathbf{x}} - \sum_{k \in s_r} w_k^{\mathrm{NA}} \mathbf{x}_k\right).$$

The resulting calibration weight w_k^C can also be viewed as a regression weight. In the LFS, a separate intercept for each province is (implicitly) included in the vector $\mathbf{x}$ so that the total number of people in each province is implicitly contained in the vector of control totals $\mathbf{t}_{\mathbf{x}}$. This is due to the fact that the 24 age-sex groups cover the entire provincial population. It can be shown that the (implicit) inclusion of the intercept in the vector of auxiliary variables leads to a calibration adjustment factor that reduces to

$$g_k^C = \mathbf{x}_k' \left(\sum_{k \in s_r} w_k^{\mathrm{NA}} \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{t}_{\mathbf{x}}.$$

This simplified expression for the calibration adjustment factor is used in the variance estimation system.

As pointed out in the introduction, calibration in the LFS is used for the following three reasons: i) to ensure consistency with Census projected estimates and with all surveys using these Census estimates; ii) to account for undercoverage to some extent and iii) 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, it is easy to see through an appropriate model that being employed or unemployed is related to the age and sex of a person.

6.4.2 Composite calibration

Composite calibration (or regression composite estimation) is essentially the same as calibration, except that some control totals are estimates from the previous month, and that the auxiliary variables associated with these estimated control totals are not known for all people $k \in s_r$ and are thus imputed. These control totals and auxiliary variables are called composite control totals and composite auxiliary variables respectively. There are 25 composite auxiliary variables for each province and they are all defined with respect to the previous month (see Section 6.5 for a complete list).

Let us now denote the vector of composite auxiliary variables for unit k by $\mathbf{z}_{t-1,k}$ and the corresponding vector of estimated control totals by $\hat{\mathbf{t}}_{\mathbf{z}}$. Since $\mathbf{z}_{t-1,k}$ is defined for the previous month (month $t-1$), the estimated control totals $\hat{\mathbf{t}}_{\mathbf{z}}$ must be computed using the previous month's data. Unfortunately, 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. To cope with this problem, imputation is performed to fill in missing values using a mixture of two imputation methods.

In the first method, we use mean imputation and obtain the modified vector:

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

where s_r^b is the subset of people $k \in s_r$ 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 method, we use the modified vector:

$$\mathbf{z}_{\bullet k}^{(2)} = \begin{cases} \mathbf{z}_{t-1,k} + (\delta_k^{-1} - 1)(\mathbf{z}_{t-1,k} - \mathbf{z}_{tk}) & \text{if } k \in s_r - s_r^b \\ \mathbf{z}_{tk} & \text{if } k \in s_r^b \end{cases},$$

where $\mathbf{z}_{tk}$ is the vector $\mathbf{z}_{t-1,k}$ defined at the current month *t* and δ_k is the probability that $k \in s_r - s_r^b$ given that $k \in s_r$. In the LFS, $\delta_k = 5/6$, for $k \in s_r$, and is replaced in the previous equation by the estimate $\hat{\delta}_k = \sum_{k \in s_r - s_r^b} w_k^{NA} / \sum_{k \in s_r} w_k^{NA}$. 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 leads to $\mathbf{z}_{\bullet k}^{(2)}$) was efficient for estimating population parameters defined as differences between two successive months.

Neither $\mathbf{z}_{\bullet k}^{(1)}$ nor $\mathbf{z}_{\bullet k}^{(2)}$ is actually used in the survey. Instead, the composite auxiliary variables are defined as

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

where α is a tuning constant that has been chosen to be equal to 2/3. This leads to a compromise between the two imputation methods. A study on the choice of α 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_k^{CC} are obtained by minimizing the distance function given in Section 6.4.1 that was used to obtain calibration weights w_k^C, except that the constraint $\sum_{k \in s_r} w_k^C \mathbf{x}_k = \mathbf{t}_{\mathbf{x}}$ is replaced by the constraint

$$\sum\nolimits_{k \in s_r} w_k^{CC} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{z}_{\bullet k} \end{pmatrix} = \begin{pmatrix} \mathbf{t}_{\mathbf{x}} \\ \hat{\mathbf{t}}_{\mathbf{z}} \end{pmatrix}.$$

Note that the composite calibration weights w_k^{CC} are still calibrated on the usual control totals $\mathbf{t}_{\mathbf{x}}$. They are given by $w_k^{CC} = w_k^{NA} g_k^{CC}$, where the composite calibration adjustment factor g_k^{CC} has the same form as g_k^C, with the exception that $\mathbf{x}_k$ and $\mathbf{t}_{\mathbf{x}}$ are replaced by $(\mathbf{x}_k', \mathbf{z}_{\bullet k}')'$ and $(\mathbf{t}_{\mathbf{x}}', \hat{\mathbf{t}}')$ respectively. 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 we assign $\mathbf{z}_{\bullet k} = \mathbf{0}$ to out-of-scope people at the previous month. The idea is to determine $\mathbf{z}_{\bullet k}$ so that $\sum_{k \in s_r} w_k^{NA} \mathbf{z}_{\bullet k}$ remains, like $\hat{\mathbf{t}}_{\mathbf{z}}$, remains an estimate of the unknown vector of control totals $\mathbf{t}_{\mathbf{z}}$, 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.

The reason for using composite calibration is to improve the efficiency of the estimates. Substantial improvement in the efficiency of the estimates for $\mathbf{z}_t$ is obtained if there is a strong month-to-month correlation

between $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$. 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 and imputation of $\mathbf{z}_{t-1}$. Overall, it was found empirically that composite calibration is beneficial in the LFS.

6.4.3 Integrated method of weighting

Since some auxiliary variables and all composite auxiliary variables are defined at the person level, the calibration weights w_k^C and the composite calibration weights w_k^{CC} are not constant within a household, unlike the subweights w_k^{NA}. This does not pose a problem as long as we are interested in estimating person-related population parameters, such as the total number of people employed in the population. In the LFS, we are also sometimes interested in estimating household-related population parameters, although to a limited extent. For example, we may be interested 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 $\mathbf{x}_k$ and $\mathbf{z}_{\bullet k}$ for a given person k by the average of $\mathbf{x}$ and $\mathbf{z}_{\bullet}$ over all members of his or her household and then computing the calibration weights or the composite calibration weights as in Section 6.4.1 or 6.4.2. 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.

6.4.4 Treatment of negative final weights

Sometimes, negative final weights occur. In this situation, composite calibration is performed again, with the difference that the subweights are replaced by the final weights when they are positive and are kept intact when the final weights are negative. If after this second round of composite calibration there are still negative final weights, then these negative weights are set equal to 1 and we accept that the composite calibration constraint will not be satisfied. This rarely occurs.

6.5 Composite auxiliary variables, 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.

<u>Labour force characteristics of previous month (no breakdown)</u>

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

<u>Labour force characteristics of previous month by age/sex groups</u>

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+*

<u>Employment of previous month by industry</u>

Employed in agriculture, 15+
Employed in construction, 15+
Employed in information, culture and recreation, 15+
Employed in utilities, 15+
Employed in manufacturing, 15+
Employed in natural resources, 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 accommodation and food services, 15+
Employed in public administration, 15+
Employed in trade, 15+
* *Employed in other services, 15+*

<u>Employment of previous month by class of worker</u>

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

Chapter 7 Variance estimation

Introduction

In a survey based on a probability sample such as the LFS, statistical inferences need to account for the sampling error. The variance measures the precision of an estimator. Due to the complexity of the estimation method and sample design, an explicit form of the variance estimator may not be available. But it is possible to obtain a variance estimate using data drawn from the sample.

The LFS uses the jackknife method to estimate the sampling variability. Section 7.1 describes the application of the jackknife method in the cross-sectional and repeated context of the LFS, while Section 7.2 presents the major steps of variance estimation in the environment of the LFS. Section 7.3 deals with the computer system that was developed for this purpose.

7.1 The jackknife method

The LFS uses the jackknife method to estimate the sampling variance of the regression composite estimator (or composite calibration estimator) used for each of the ten provinces, and of the generalized regression estimator used for each of the three territories. The jackknife technique is a re-sampling procedure which requires that strata and replicates be defined for variance estimation purposes. In the 2004 design of the LFS, for all provinces and in many cases for the three territories as well, the variance strata and replicates are respectively identical to the design strata and primary sampling units (PSUs). This is in accordance with Särndal, Swensson and Wretman (1992), Remark 11.5.2, "In multistage sampling, the jackknife technique is usually applied at the PSU level". Due to the selection of only one PSU in a few strata in the territories, stratum collapsing is sometimes required in order to have at least two replicates per stratum. When two or more strata are collapsed, then each original single-PSU stratum becomes a replicate for variance estimation purposes.

In each province and territory denoted by *p*, for variance stratum *ph* and a given replicate *phi*, a jackknife replicate $\hat{Y}_{p(\text{phi})}$ of a monthly total estimator $\hat{Y}_p$ is obtained as follows. The estimation procedure is repeated at the province or territory level, after omitting the sample records for replicate *phi* and multiplying the subweights of the remaining records in variance stratum *ph* by the factor $A_{ph}/(A_{ph}-1)$, where A_{ph} is the number of replicates in stratum *ph*, in order to account for the omitted records. Similarly, at the Canada level, a jackknife replicate $\hat{Y}_{C(\text{phi})}$ of a monthly total estimator $\hat{Y}_C$ is obtained by repeating the calibration procedure at the Canada level, after omitting the sample records for replicate *phi* and, as before, adjusting the subweights of the remaining records in variance stratum *ph*.

An estimator of the sampling variance of a provincial or territorial monthly total estimator $\hat{Y}_p$ is given by

$$\hat{V}(\hat{Y}_p) = \sum_{h=1}^{H_p} [(A_{ph}-1)/A_{ph}] \sum_{i=1}^{A_{ph}} [\hat{Y}_{p(\text{phi})} - \hat{Y}_p]^2, \qquad (1)$$

while an estimator of the sampling variance of a Canada level estimate $\hat{Y}_C$ is given by

$$\hat{V}(\hat{Y}_C) = \sum_{p} \sum_{h=1}^{H_p} [(A_{ph}-1)/A_{ph}] \sum_{i=1}^{A_{ph}} [\hat{Y}_{C(\text{phi})} - \hat{Y}_C]^2, \qquad (2)$$

where H_p is the number of variance strata in province or territory *p*. The variance of monthly totals is additive over provinces since sampling is independent from one province or territory to another and estimation procedures are applied to each province or territory individually, so that $\hat{Y}_{C(\text{phi})} - \hat{Y}_C = \hat{Y}_{p(\text{phi})} - \hat{Y}_p$; however, this is not true for many other statistics such as rates.

Singh, Kennedy and Wu (2001) state conditions necessary for the validity of the jackknife method for cross-sectional or repeated surveys. Replicate level estimates must have identical mean and variance, and replicate selection must be, or must be assumed to be with replacement. If replicates are selected without replacement, the jackknife variance estimator becomes conservative if the covariance between replicates is negative. For repeated surveys, if replicates are common (or connected) over time then this must be accounted for in the jackknife procedure.

As mentioned in Section 3.4, rotation of sampling units occurs at each stage of the multi-stage design. Households within PSUs rotate out of the sample every six months, while the PSUs themselves may rotate out after a few years. PSUs that rotate into the sample are assigned the same replicate number as the PSUs they are replacing. Hence replicate vectors, representing many months of data, can be created and may be used to estimate, for example, the variance of a three month moving average.

The number of variance strata and replicates may be different from one month to the next. This may be due to an area being inaccessible because of weather conditions, the introduction or removal of PSUs following a sample size increase or decrease, or a relatively small sample size PSU having no real or

imputed respondent data in some months (all dwellings are vacant, seasonal, under construction, or occupied by persons not to be interviewed). The set of replicate vectors over, say, a period of six months, is the union of the six monthly sets of replicates. It is therefore possible, if the monthly sets of replicates are not identical, that the union contains variance strata and/or replicates not present in a given month. The estimate at the province level would then be used instead for the missing replicates.

It follows from the previous description of the jackknife variance estimator that each variance stratum needs to contain at least two replicates. Only one PSU is selected, however, from a few remote strata in the 2004 design of the LFS, where collection cost is substantial and the total number of PSUs in each stratum is small. In order to meet the requirement of having at least two replicates selected in each stratum, one of two alternative strategies may be used: i) collapse the one-PSU strata by province and consider each stratum as a replicate, or ii) assume two PSUs are selected from each stratum but the second PSU represents a pseudo-population and therefore yields zero estimates. A simulation study using data from the 2001 Census of Population showed there is little difference in estimates and variance estimates between the two strategies and, as a result, the second strategy, which is already built in the LFS variance estimation system, was adopted.

Provided that replicates are common or connected over time, the jackknife technique is applicable to a linear or non-linear function of monthly totals, such as a monthly rate, a month-to-month difference in rates or an annual average. The jackknife variance estimate of a month-to-month difference in totals for a given characteristic *Y,* for example, may be computed from Equation (1) after replacing $\hat{Y}_p$ and $\hat{Y}_{p(\text{phi})}$, respectively, by $\hat{Y}_{p1} - \hat{Y}_{p2}$ and $\hat{Y}_{p1(\text{phi})} - \hat{Y}_{p2(\text{phi})}$, where the numeric subscripts refer to the month and phi refers to replicates over time. More generally, the variance estimate of a linear or non-linear function g of monthly totals for a province or territory p is obtained from Equation (1) after replacing $\hat{Y}_p$ and $\hat{Y}_{p(\text{phi})}$, by $g(\hat{\mathbf{Y}}_{p1}, \hat{\mathbf{Y}}_{p2}, ..., \hat{\mathbf{Y}}_{pM})$ and $g(\hat{\mathbf{Y}}_{p1(\text{phi})}, \hat{\mathbf{Y}}_{p2(\text{phi})}, ..., \hat{\mathbf{Y}}_{pM(\text{phi})})$, respectively, where $\hat{\mathbf{Y}}_{pm}$ represents a vector of estimates for month m and province or territory p, and $\hat{\mathbf{Y}}_{pm(\text{phi})}$ represents an associated vector of jackknife replicates. Equation (2) is used in a similar way to estimate the variance of a linear or non-linear function of monthly totals at the Canada level. Equation (2) is well suited for the estimation of the sampling variance of rates for more than one province as, for example, unemployment rates at the Canada level, which are estimated as combined ratios in the LFS.

As a re-sampling method for variance estimation, the jackknife procedure is computationally intensive but straightforward to implement. In the context of the LFS, the procedure remains relatively simple to apply. A comparison of commonly used re-sampling methods may be found in Rao (2005) and Särndal, Swensson and Wretman (1992).

7.2 Variance estimation for the regression composite and generalized regression estimators

In order to estimate the sampling variability of an estimator, each of the steps leading to the computation of the final weights should, in theory, be repeated for each jackknife replicate. In the case of the LFS however, only the calibration procedure is repeated. The cluster, stabilization and nonresponse adjustment weights, which together yield the subweights used for calibration, are computed only once. Calibration for the generalized regression and regression composite estimators is described, respectively, in Sections 6.4.1 and 6.4.2. As we will see in the next paragraph, the variance estimation for the regression composite estimator, which utilizes data from two consecutive months, is more elaborate than that for the generalized regression estimator.

There are three steps involved in obtaining composite final weights at the province level:

i) Using the previous month's tabulation file with composite final weights, obtain the previous month's estimates of key labour market characteristics, adjusted to be consistent with the current month's population totals. This step yields the regression composite control totals.

ii) Merge records from the current and previous months' tabulation files for the five rotation groups in common between the two files, in order to impute composite auxiliary variables for records missing in the previous month's file. Append the resulting file with records in the birth rotation group from the current month's tabulation file. Composite auxiliary variables missing for the birth rotation will be imputed by the mean, as described in Section 6.4.2.

iii) Calibrate the current month's subweights from the file obtained in step ii) up to the current month's demographic and regression composite control totals using the generalized regression calibration

procedure. This step yields the current month's composite final weights.

For variance estimation, steps i) and iii) are repeated as part of the calibration process for each jackknife replicate, while step ii) is undertaken only once, at the province level. It may be noted also that the occurrence of negative weights, as discussed in Section 6.4.4, is not accounted for during variance estimation.

As mentioned in the previous section, the replicates within variance strata are common or connected over time in order to estimate, for example, the variance of an annual average. For the purpose of estimating the variance of the composite estimator during the sample phase-in period when a new design is introduced, the replicates of the 1984 and 1994 sample designs, as well as those of the 1994 and 2004 designs, were linked. A cluster overlap file between consecutive sample designs was used for the linkage. Although such a linkage between consecutive sample designs can only be approximate, it provided for a reasonable transition in variance estimates during the sample phase-in period.

7.3 The LFS variance estimation system

The LFS variance estimation system was developed to meet the following objectives: using the jackknife method, i) obtain estimates and variance estimates of user defined linear or non-linear functions of monthly totals, such as, for example, a three-month moving average of total employment by province, or a difference in unemployment rates for two domains in the same month and, ii) compute design effects of monthly totals, for both of the regression composite and generalized regression methods of estimation. Variance estimates may currently be obtained for the 1984, 1994, and 2004 sample designs as well as for the sample phase-in period between consecutive designs. The program is applicable to the LFS as well as to other multistage stratified surveys which use the same type of estimator.

Chapter 8 Data quality

Introduction

The LFS estimates, like those produced from any sample survey, can contain sampling and non-sampling errors. As a result, to correctly interpret the estimates of this survey, it is important to be aware of their quality.

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 probably differ from those we would get if we conducted a complete census of this population in the same conditions. An 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.

A non-sampling error, as its name indicates, has nothing to do with 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. We can also associate non-sampling error 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 presented in two sections below: sampling errors and non-sampling errors.

8.1 Sampling errors

The repercussions of sampling errors on survey estimates depend on several factors (see Section 2.2 for a definition of sampling error). The most obvious is sample size. If all other factors are constant, the sampling error generally decreases as the sample size increases. Sampling errors also depend on factors such as variability of the population, the estimation method and the sample design.

For a sample of a given size, the sampling error is linked to various characteristics of the sample design, such as the stratification method used, the sample distribution, the choice of sampling units, and the selection method used at each stage of a multi-stage sample design.

In addition, the estimation method used plays an important role for a given sample design. The LFS has made a major change to its estimation method. This new method, called composite estimation (see Chapter 6), significantly reduces sampling errors.

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.

One of the main characteristics of a probability sample, such as the one used in the LFS, is that the sampling variance can be estimated through the sample. Chapter 7 describes the method currently used to estimate the variance of the estimates produced by the LFS.

Three measurements are derived from the sampling variance: standard error, the coefficient of variation (CV) and the design effect (see Section 2.2). The standard error 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.

To highlight the links between the different measures of accuracy, let us look at the following example. In March 2005, the unemployment rate of the Canadian population 15 years of age and up was 7.4% and the standard error was estimated at 0.0013. Consequently, the coefficient of variation was 1.8% (0.0013/0.074). The 95% confidence interval was between 7.14% and 7.66%, or 0.074±0.0026. This means that if the selection process were repeated many times, 95% of the confidence intervals we observed would contain the value we would get using a census.

Given their stability, the CVs included in the monthly LFS publication are not updated every month. Instead,

we provide an estimate of the CV that corresponds to the average of the CVs from the previous year. These estimates are updated twice a year (January to June and July to December). The table below provides the CVs observed for the monthly employment and unemployment estimates

Table 8.1 Monthly coefficients of variation (CV) observed, 2005

Province	Employed	Unemployed
	%	
Newfoundland and Labrador	1.6	5.7
Prince Edward Island	1.4	7.5
Nova Scotia	0.9	5.6
New Brunswick	1.0	5.5
Quebec	0.6	4.0
Ontario	0.4	3.2
Manitoba	0.7	7.0
Saskatchewan	0.8	7.4
Alberta	0.6	5.3
British Columbia	0.7	5.3
Canada	0.3	1.9

Thanks to the data collected during the LFS, we can produce thousands of estimates of population characteristics, as well as change estimates from one month to the next, annual estimates, and national, provincial and subprovincial estimates. Due to space limitations in current and special publications, we cannot include the CVs of all the survey estimates published. However, there are tables that present the approximate CVs for different estimate groups. They are available in the Labour Force Historical Review (Statistics Canada Catalogue No. 71F0004X).

The change estimates from one month to the next have become more important over time. In this respect, the monthly LFS newsletter now provides the standard errors (SEs) for the provincial and national differences for employed and unemployed. These figures for 2005 are provided in the table below.

Table 8.2 Standard error (SE) of the variation from one month to the next, *Employed* and *Unemployed*

Province	Employed	Unemployed
	thousands	
Newfoundland and Labrador	3	2
Prince Edward Island	1	1
Nova Scotia	4	3
New Brunswick	3	2
Quebec	18	14
Ontario	20	15
Manitoba	4	3
Saskatchewan	3	2
Alberta	9	6
British Columbia	12	9
Canada	32	24

We can use the design effect as an indicator of the deterioration of the sample design over time. In the LFS, we calculate two types of design effects and each one depends on the data used to establish it. We determine the *unadjusted design effect* using uncalibrated weights, meaning without an adjustment that takes the population counts into consideration. We calculate the *adjusted design effect* using the final weights. As a result, the unadjusted design effect indicates the effectiveness of the sample design, while the adjusted effect provides a general evaluation of the strategy adopted by combining all the characteristics of the survey plan (stratification, multi-stage sampling, poststratification and estimation). The smaller the effect, the more effective the design with regard to sampling variance. Monitoring the design effect helps to evaluate qualitative changes made to the design over time. We stress 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 poststratification.

In the LFS, we use the unadjusted design effect together with other information to identify regions where the sample design has lost a significant portion of its effectiveness over time. In some cases, we must do a mini-redesign in these regions to remedy this problem. The table below presents some values representing the adjusted and unadjusted design effects for the characteristics employment and unemployment at the national and provincial levels, based on survey data from January to June 2004.

Table 8.3 Design effects, Employed and Unemployed, 2004

Province	Employed		Unemployed	
	Adjusted	Unadjusted	Adjusted	Unadjusted
Newfoundland and Labrador	0.68	2.44	1.10	1.23
Prince Edward Island	0.42	1.86	1.27	1.21
Nova Scotia	0.40	2.81	1.07	1.06
New Brunswick	0.50	3.06	1.28	1.35
Quebec	0.35	2.77	1.02	1.11
Ontario	0.32	6.21	1.03	1.19
Manitoba	0.23	3.73	0.97	1.03
Saskatchewan	0.31	3.94	0.97	0.93
Alberta	0.28	9.56	1.03	1.26
British Columbia	0.33	3.08	1.01	1.10
Canada	0.33	4.87	1.03	1.16

8.2 Non-sampling errors

Non-sampling errors can occur in any step of the survey and are generally caused by human error, such as a lack of attention, and poor understanding or interpretation. The impact on the estimates can be seen in the bias and/or variability of the estimates. The net effect of non-sampling error variance may be minor if there are many observations or for large domains. However, its

effect can be large for small domains or when the characteristics being studied are rare or associated with sensitive issues.

In addition, the net effect of non-sampling bias tends to be additive. This bias can be due to interviewer's training or attitude, poor questionnaire design, or the imputation method used to resolve nonresponse. All these factors can contribute to an accumulation of errors in one direction or the other.

Non-sampling variance and/or bias can come from different sources. Below, we will look at coverage, non-response, vacancy, response, processing and field activities.

8.2.1 Coverage errors

Coverage errors can arise during several steps of the survey process, such as when the survey frame is being created, when the dwellings and/or persons to include in the survey are identified, or when data are collected and processed. In the LFS, the indicator used to measure coverage error is called the slippage rate. This rate is the relative difference between the population size estimates produced from the pre-calibration weights and the most recent population estimates from the census. The population 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, we observe undercoverage, indicated by a positive slippage rate. To reduce the resulting bias as much as possible, we adjust the weight for each respondent based on control totals from independent sources (see Chapter 6).

Omitting dwellings or persons from the target population, or in other words, the presence of undercoverage in the LFS can introduce non-sampling errors. By dwellings, we mean any habitable construction that meets certain criteria. The persons who live in a dwelling comprise the household. 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 classified as vacant in error. 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, slippage biases the unemployment estimates downward. Lastly, as mentioned earlier, the population estimates also play a role in slippage.

Other factors that can contribute to slippage in the LFS were identified. For example, the population grows between redesigns, generally in specific places and unevenly. The sample can over- or underestimate this growth or accurately account for it. Furthermore, the adjustment to account for nonresponse (see Chapter 5) can also influence slippage. In reality, if non-respondent households have fewer members and are represented in the sample by large households, this can affect the slippage rate.

Every month, the slippage rates are thoroughly analyzed. They are produced monthly for the regions and 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+). They are also produced for each territory, but with no breakdown by age-sex group. In the last LFS redesign, we produced revised slippage rate series that use population estimates based on the 2001 Census. The table below provides the average slippage rates for the 2005 calendar year.

Table 8.4 Average slippage rates - Canada by age group and province, 2005

Canada	%
All ages	10.3
15 to 19	8.3
20 to 24	18.3
25 to 29	19.8
30 to 39	14.6
40 to 54	8.8
55+	4.7
Newfoundland and Labrador	8.4
Prince Edward Island	7.0
Nova Scotia	8.0
New Brunswick	9.0
Quebec	7.9
Ontario	10.8
Manitoba	6.1
Saskatchewan	7.1
Alberta	13.2
British Columbia	14.2

Finally, we periodically produce estimates of the number of households by household size. These estimates provide another point of view on slippage.

All these indicators serve to detect potential problems with the sample coverage and to assist in taking any necessary action. To remedy or slow its progression, we

can, for example, consider creating exercises for interviewers to increase their knowledge of the household composition rules, 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.

Slippage will always be monitored closely since it can introduce a bias into the estimates. Moreover, despite applying an estimation method to correct the slippage, we can expect a certain bias to persist in the estimate, other than the usual estimation bias, since the characteristics of the omitted persons and dwellings can differ from those of the persons included in the sample.

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 persons who are not part of the target population, *e.g.* members of the Canadian Armed Forces;
- vacant dwellings: unoccupied or seasonal dwellings or dwellings under construction;
- non-existent dwellings: demolished dwellings, dwellings turned into business locations, mobile homes moved, abandoned dwellings, or a dwellings entered by mistake.

When a dwelling is identified as eligible for the survey, we cannot always do an interview for the following reasons:

- household nonresponse: no one at home, temporary absence, interview impossible (inclement weather, unusual circumstances in the household, *etc.*), technical problems, or refusal.

The importance of the bias due to nonresponse is usually not known, but we do know that 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, we try to maintain the response rate as high as possible during collection.

Since 1993, the LFS has been following the Statistics Canada standards and guidelines for declaring non-response rates. Every month, the weighted and unweighted nonresponse rates are sent to the Statistics Canada Central Nonresponse Database, which is mandated to compile the longitudinal data for a number of regular surveys. This database requires the non-response rates at the collection and estimation steps.

The table below presents the average nonresponse rates as well as the minimum and maximum rates attained for 2005. In the LFS, the maximum non-response is usually attained in July, given the high percentage of persons who are not at home, while the minimum is attained in October.

Table 8.5 Nonresponse rates (unweighted), Canada and the Provinces, 2005

Province	Average	Maximum	Minimum
		%	
Newfoundland and Labrador	4.2	5.4	3.0
Prince Edward Island	3.5	4.8	2.4
Nova Scotia	6.3	7.3	4.6
New Brunswick	4.6	5.4	3.1
Quebec	5.4	6.6	3.7
Ontario	4.8	5.7	3.7
Manitoba	3.6	5.4	2.1
Saskatchewan	3.6	4.6	2.4
Alberta	4.9	6.3	3.1
British Columbia	5.7	6.7	4.5
Canada	4.9	5.5	3.8

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

Since 1999, several factors have disrupted the LFS nonresponse rate series. (For more details see LFS 2005). First, the introduction of a new computer-assisted interview (CAI) system in the regional offices in Fall 1999 brought about technical difficulties, which caused an increase in nonresponse at the beginning of its implementation. The computer-assisted telephone interview (CATI) system introduced in September 2000 for subsequent interviews also affected the nonresponse series. During the same period, the arrival of new laptop computers improved the application's performance, and the series returned to its regular level, which was that just before the new CAI system was introduced. Moreover, the new sampling design introduced in November 2004 also affected the series, in that it required the hiring of new interviewers who have a tendency to obtain slightly higher nonresponse rates in their first six months with the LFS. Lastly, during the same period, the introduction of the Telephone First Contact methodology for new dwellings in the sample generated a new type of nonresponse: cases transferred to the field. For a historical background of the

nonresponse issues with the LFS, please refer to the article by Sheridan *et al*. (1996).

Since Telephone First Contact (TFC) was implemented (see Chapter 4), each month approximately one-third of the new dwellings are contacted by telephone to respond to the LFS. The LFS managers were concerned with the possible effects of these changes on refusals. As a result, special attention was paid to this component and a number of tables were produced and thoroughly analyzed. We note a higher nonresponse rate for cases with "Telephone First Contact" compared to those with "First Contact by a Personal Visit". However, this higher nonresponse rate seems primarily due to an increase in cases with no contact for the households in the TFC group.

Refusal rates for the LFS are usually very low, with monthly Canadian rates varying between 1% and 2%. The provincial refusal rates are usually similar, but can dip as low as 0.5% or climb as high as 3%. The collection system makes it possible to get more information on the reason for refusal, and therefore we can keep track of the changes in respondents' attitudes toward the survey over time.

8.2.3 Vacancy

Dwellings correctly identified as being vacant or non-existent do not introduce a bias into the LFS estimates. However, the estimation variance is higher because the sample contains fewer households. The LFS interviewers return to the vacant dwellings every month to interview the persons targeted by the survey who could have moved there since the previous survey. Non-existent dwellings are simply removed from the survey frame. Special attention must be given when determining the vacant dwellings since they have a direct influence on two other indicators. If a dwelling is coded as vacant but its occupants are temporarily absent, the nonresponse rate produced for the LFS will be underestimated somewhat. 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. In the LFS, the "Program to verify vacant dwellings" was established to obtain information on this error.

The table below presents the average vacancy rates and the minimum and maximum values for 2005 at the provincial and national levels.

Table 8.6 Vacancy rate (unweighted), Canada and the Provinces, 2005

Province	Average	Maximum	Minimum
		%	
Newfoundland and Labrador	15.4	14.9	16.4
Prince Edward Island	20.5	18.6	23.0
Nova Scotia	16.8	15.2	18.7
New Brunswick	14.1	13.5	15.2
Quebec	14.0	11.9	15.8
Ontario	10.8	10.0	11.3
Manitoba	17.1	16.4	17.7
Saskatchewan	14.7	12.5	15.5
Alberta	8.7	8.1	9.8
British Columbia	9.5	8.7	9.8
Canada	13.0	12.2	13.5

Generally speaking, the vacancy rate is relatively stable, with an upward trend the further we are from the last redesign since the survey frame is less up-to-date. After each redesign, the vacancy rate decreases. For this quality indicator, some provinces set themselves apart from others with much higher or lower rates.

8.2.4 Response error

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, and the general survey conditions. They can occur when the information is provided, received or entered into the computer. However, with the computerized collection method, we can 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, have forgotten or altered the facts for personal reasons. In addition, interviewers can have a tendency to explain responses or interpret them differently. As in the other error categories, response errors can have a variance and a bias.

The proxy responses we get from one household member when we collect information about another household member can also lead to response errors.

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 attains its highest level for the sixth of the sample in its first interview. We can 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, *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 among the rotation groups, but whose tangible effect is minor.

8.2.5 Processing errors

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

Using a computerized collection method helps to prevent skip errors in the questionnaire, since it is now the application that determines the next question to ask, given the previously entered responses. Similarly, certain verification rules are integrated into the collection system to detect and correct discrepancies at the time of the interview.

During processing, we assign a validation code to each response entered by the interviewer. This code indicates whether the respondent did not know the response to the question, whether the response entered was rejected when the verification rules were applied, whether the respondent refused to respond, or whether the response entered contained superfluous information that must be deleted. The latter code applies only to data from central office computers. To improve error control at the validation stage, the distribution of these codes is occasionally reviewed.

The field control module also provides the discrepancy rate from the form control. A discrepancy is defined as any entry deleted, modified or added to a blank field after undergoing certain controls to verify its validity. The discrepancy rate at the control automatically represents the percentage of discrepancies on a questionnaire compared to the total number of entries on the questionnaire. The rates are calculated based on various verification procedures applied at the central LFS office. The corresponding discrepancy rates at control are 0.1% and 1%, respectively.

The variables "occupation" and "industry" are coded automatically and manually at the central office. In the first month of interviews or when any change is made to these two variables, 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 automatically coding the descriptive information collected for the variables "occupation" and "industry" based on the standard classification for these variables, SOC and NAICS. The records that could not be coded by the automatic system are coded manually by a team of coders at the LFS. Approximately 14,000 records are coded manually every month. In order to control the quality of the manual coding, a statistical quality control plan is applied monthly. The three measurements used to determine the effectiveness of this control process are the verification rate**,** or the percentage of records verified out of all the records submitted for control, the incoming error rate (IER), an estimate of the percentage of records that contain errors before undergoing quality control, and the outgoing error rate (OER), an estimate of the percentage of records that still contain coding errors following quality control. In 2005, the average value of these three measurements was 19.6% for the verification rate, 7.9% for the IER, and 4.8% for the OER.

The imputation rate is also a quality indicator with regard to data processing. The LFS occasionally produces imputation rates, broken down by imputation method, by questionnaire and by question. This indicator 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, we thoroughly assess the result of these activities, analyze the different diagnoses automatically produced by the system, and do a comparison with other data sources.

8.2.6 Errors during collection

The collection application produces files that contain a host of information on what goes on in the field. Using these files, we can produce a myriad of quality indicators on the interviewers' activities in the field. The LFS regularly analyzes the calls and visits made by the interviewers as well as the duration of the interviews. For example, we can determine the average time spent per personal or telephone interview, the number of attempts to reach a respondent, the best times (time and day) to conduct an interview, *etc*. We can also check whether the interviewers strictly follow the collection procedures.

The application therefore provides more information on the work carried out in the field and allows us to take action in questionable cases (*e.g.* interviews done in less than one minute), and therefore to reduce certain errors that can creep in. 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. These data also allow us to find out more about the editing rules applied during the interview. For example, we can find out how many editing rules were exceeded after confirmation with the respondent, how many times an editing rule was applied, and how many times an observation fell outside the rules.

All these measures will help to better understand what goes on in the field and when interpreting the results.

8.3 LFS data quality reports

Some reports are made available to LFS data users at the central office in Ottawa or in the regions. These documents are also consulted regularly by the members of the LFS Data Quality Committee to maintain the quality of the survey. They contain a wide range of quality indicators at different geographical levels and for shorter or longer periods.

LFS Monthly Survey Operations Report. Every month, the LFS data quality unit produces a report on the quality of the survey data for the current month. Its main purpose is to monitor the quality of field operations, which is why most of the quality indicators are presented by regional office. Certain series were also presented for a period of 26 months to get a better idea of the seasonal and monthly changes compared to the previous year. The report contains the following quality indicators: nonresponse rate (by regional office, component, number of months in the survey, and urban/rural), vacancy rate (by regional office), slippage rate (by province and age-sex group), sample size, the number of technical problems and the number of temporary dockets (interviewer-selected dwellings).

Variance tables. The monthly variance tables contain the estimates, CVs, variances and design effects of the main LFS characteristics at different geographical levels. In addition to these indicators, the average household size and slippage rates are given at more detailed levels than those presented in the LFS Operations Report.

Quality Report. The LFS Quality Report is produced every year. It presents an in-depth review of the LFS quality measures and an analysis of the quality measures over a 30-month period to detect any trends or the effects of certain changes made to the activities or the sample design. This document contains several quality indicators presented in the Operations Report, but here we look at the provincial rather than the regional level. In contrast to the Operations Report, the Quality Report includes analysis of the different tables and graphs. It also contains a special chapter on a particular subject of interest.

Special reports. In addition to the regular reports produced to ensure and control LFS data quality, special reports are written on occasion. For example, one recent study looked at the potential impact of reducing the collection period from ten days to nine days on response rates and estimates.

8.4 Quality assurance programs

Over the years, the LFS has created a number of programs to ensure the quality of the data it publishes. Seven of these programs are presented below.

Recruitment. Before hiring candidates for interviewer positions, we evaluate their skills and ability to properly complete the survey documents. Even before their training begins, we send them a copy of several documents which describe the work of Statistics Canada interviewers (responsibilities, techniques and required competencies) and the Statistics Canada organization.

Training. The initial training period for interviewers is two months. It begins with a three-day course in the classroom, during which we show the new interviewers how to use the computer equipment and how to complete the survey forms and administrative documents. They also do practical exercises, simulate interviews and learn interview techniques.

Then, the interviewers receive two days of on-the-job training during their first survey week and one or two days during the second week, if necessary. A senior interviewer accompanies and observes them, and explains and demonstrates how to conduct the interviews. All interviewers also attend special group training and retraining sessions at least once a year.

The interviewers' work is evaluated as part of other programs, which we will describe later. Based on each interviewer's performance, we determine whether self-training courses or review exercises are necessary to clarify some points or resolve weaknesses.

Observation. The purpose of the observation program is to reduce as much as possible the errors that interviewers could make by giving the senior interviewer the opportunity to observe the interviewers under his/her supervision, evaluate their performance and identify any problems. Each interviewer is observed at least once every 24 months. The regional office decides who will be observed and when, so no one can predict the order

of the observations. Outside this program, the senior interviewer can observe one of his/her interviewers if he/she suspects a specific problem. The senior interviewer will accompany the interviewer for an entire day and see how the personal and telephone interviews are conducted. On the second day, the senior interviewer checks the cluster lists, then sends the observation results to the regional office. He/she also writes periodical reports for the central office. The senior interviewer sends the interviewer the result of his/her performance as soon as possible following the observation.

Feedback on performance. Every month, a report is written about each interviewer's performance. These reports include the costs, rejection rates at verification, and the response rates. The senior interviewers are in regular contact with their interviewers and will inform them of the results of the various performance indicators.

Verification of dwellings coded as vacant. The program to verify dwellings coded as vacant aims at monitoring the work done by the interviewers in the field. For each interviewer, a sample of dwellings coded as vacant is selected at least once every 24 months. The senior interviewer returns to these dwellings to check whether they are indeed vacant and the dwellings are recoded (vacant or other). Once the results are in, the interviewer receives additional training, if deemed necessary. This information is also used to measure how many households coded out of scope for the survey contributed to slippage, although it is very difficult to extrapolate to the full sample since the choice of verified dwellings is up to the regional offices.

Validation program. The validation program was designed to monitor the performance of interviewers and to give interviewers constructive feedback in the form of additional training based on the weaknesses identified. Interviewers are validated randomly so that each one is selected twice a year. Approximately 2% of households are included in this program every month (except in April and December, when the program does not run). The week following the survey week, the senior interviewers re-contact the persons who provided information during the survey week for the sample submitted to this program. They ask them questions such as address confirmation, if they remember taking part in the survey, when the interview was, the interviewer's attitude, etc. They also take this opportunity to thank the respondents for their participation and their time.

PSU performance control. The yield of the PSUs is monitored monthly to detect any differences between the number of dwellings surveyed in the field and the number of dwellings used in the creation of the sample design. The sample design uses a number derived from the enumerations performed using data from the previous census. As a result, any significant discrepancy, such as 50% (positive or negative), between the census and the derivation 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.

8.5 LFS committees

The LFS needs several coordination groups to see that the survey runs smoothly. Some LFS committees are permanent, while others are only active during the redesign. In the last redesign, the Redesign Steering Committee was the main high-level committee. It had the mandate to control the redesign in its entirety, i.e., to ensure that activities having to do with the sample redesign ran smoothly.

We describe three permanent committees below. Their mandate is to look after permanent operations and to evaluate the survey on a regular basis.

Operations Committee. The mandate of this committee is to review the activities that occurred during each survey month and the circumstances surrounding the conduct of the survey, to ensure that the operations run smoothly, to examine proposed changes and recommend that they be adopted. This is to ensure that the survey continues to achieve its objectives. The Operations Committee is chaired by a member of Labour Statistics Division and meets every two weeks.

Population Estimates Steering Committee. The mandate of this committee is to review the postcensal population estimates required by the LFS. It also evaluates the data sources used and the methods applied to obtain the estimates at different geographical levels, and initiates a number of research projects on the subject. This

committee is chaired by a member of the Labour and Household Survey Analysis Division.

Data Quality Committee. The committee, which was officially created in the spring of 1972, was at that time responsible for publishing the quality of the LFS data and its supplements. Since then, its mandate has expanded somewhat and now consists of examining and evaluating the quality of LFS data on a monthly basis, proposing and reviewing research and development projects aimed at tweaking methods that can affect data quality, and monitoring research and development in this field. 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 aspect likely to improve quality. The committee members have some documents available to them to accomplish their task. By closely following the evolution of the indicators, the committee can intervene immediately with those in charge of the LFS activities in question to control the quality of the monthly data. It 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*.

Chapter 9 LFS frame for other surveys

Introduction

Many household surveys use the Labour Force Survey frame and sample to collect information. These surveys can generally be described as either Special, Supplementary or Rotate-out surveys. Special surveys use the LFS frame to select a separate sample of households, usually in PSUs that are active in the LFS. Supplementary surveys interview households that have just been selected for the LFS. Rotate-out surveys are similar to supplementary surveys but contact the household after LFS has rotated it out. Some of these surveys also use active LFS households, but conduct the interview at a later date. Special and supplementary surveys are important parts of the Statistics Canada household surveys program and are often sponsored by other government departments.

Special surveys, although in separate households, can often share interviewer resources with the LFS since they are usually in the same PSUs. Supplementary and rotate-out surveys can also take advantage of data collected by the LFS to screen respondents. Note that supplementary surveys can be divided into two types: a dependent supplement that uses LFS households while they are still being interviewed for the LFS and an independent supplement that breaks off from the LFS to be interviewed at a separate time, or to allow more time than the LFS would for data collection. When the LFS rotates a household out, that household is still eligible for rotate-out surveys for up to 2 years. Use of the LFS frame and sample in this way results in substantial cost savings for these surveys.

Special surveys reserve a set of random starts to select dwellings for their exclusive use. Based on the desired allocation, each stratum in the LFS may have 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 the special survey. Samples for the special surveys are selected using these starts. This strategy reduces the respondent burden since the same dwelling is not selected for both the LFS and the special surveys.

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. Supplements must comply with Statistics Canada criteria for the reliability of data and the confidentiality of responses. 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 from 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 them to the required number of households, as in the LFS stabilization program. Selection can also take place within households by either random sampling or by screening for individuals with specific characteristics.

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 documents or through special questions.

The following list shows some of the surveys using the LFS frame or sample in 2005.

Survey	Data collection period -2005
Supplementary surveys	
International Travel Survey - Airports (ITS)	April to May
Future to Discover Project (FTDP)	April to June
Canadian Internet Use Survey (CIUS)	November
Travel Survey of Residents of Canada (TSRC)	January to December (monthly)
The Communities Survey	February to June
Residential Telephone Service Survey	December
The Survey of Northern Children	February to July
Rotate-out surveys	
Survey of Labour and Income Dynamics (SLID)	January to March
National Longitudinal Survey of Children and Youth (NLSCY)	September to June
Longitudinal Survey of Immigrants to Canada (LSIC)	January to June
Special surveys	
Canadian Community Health Survey (CCHS)	January to November (monthly)
Survey of Household Spending (SHS)	January to March
Survey of Financial Security (SFS)	May to June
National Population Health Survey (NPHS)	February, June, August, November

9.1 Examples of special and supplementary surveys

The Canadian Community Health Survey (CCHS) is a continuous survey that collects monthly data on a variety of health topics. The sample from the LFS frame is supplemented using a telephone list frame and, in a few Health Regions, by random digit dialing (RDD).

The Survey of Household Spending (SHS) is an annual household survey that, in its current configuration, was introduced as part of Statistics Canada's Program to Improve Provincial Estimates (PIPES). The Survey of Household Spending also continues to be used in its traditional role as a source of information for computing the Consumer Price Index. The SHS is a special survey, *i.e.*, it selects households in PSUs containing an LFS sample, but the SHS households are not interviewed by the LFS.

The Survey of Financial Security (SFS) is conducted occasionally to collect information on the net worth (wealth) of Canadian families, that is, the value of their assets less their debts. The LFS frame sample is also augmented with addresses in high-income geographical areas.

The Survey of Labour and Income Dynamics (SLID) was first introduced in 1993 to study the processes that influence the economic life of Canadians. The survey is used to investigate movements into and out of low-income status, labour market transitions and the relationship between family dynamics and economic well-being. Panels are selected every three years, while each panel is in the survey for six years. Each panel initially consists of households that were recently interviewed by the LFS (rotate-out). Like other longitudinal surveys, SLID follows sampled *individuals* over time.

The National Longitudinal Survey of Children and Youth (NLSCY), which began in 1994, tracks a sample of children over many years to monitor their development from infancy to adulthood. It is a complex survey which began with LFS-based households to obtain a sample of children. In addition to the original cohort of 0-11 year-olds sampled at the first cycle, the NLSCY also selects a new sample of 0-11 year-olds at every subsequent cycle. Since only 3% of the LFS households have children born in a given year, the NLSCY uses up to 20 rotation groups in the smaller provinces.

References

Alavi, A., and Beaumont, J.-F. (2004). Nonresponse Adjustment Plans for the Labour Force Survey. Paper presented at Statistics Canada's Advisory Committee on Statistical Methods, May 2004.

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.

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

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

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.

Brodeur, M., Montigny, G. and Bérard, H. (1995). Challenges in developing the National Longitudinal Survey of Children. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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.

Chen, E.J., Gambino, J., Laniel, N. and Lindeyer, J. (1994). Design and estimation issues for income in the redesign of the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Chen, E.J., Lindeyer, J and Laflamme, G. (2004). Issues with high cost areas in the Canadian Labour Force Survey sample redesign. *Proceedings of the Survey Methods Section,* Statistical Society of Canada.

Cillis, P. (2004a). LFS Mapper Functional Specification, Geography Division, internal document.

Cillis, P. (2004b). LFS Mapper User Guide. Geography Division, internal Document.

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

Deville, J.-C., and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.

Drew, J.D., Bélanger, Y. and Foy, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology*, 11, 95-110.

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.

Dufour, J., Simard, M., Allard, B. and Ray, G. (1996). Redesign of the Labour Force Survey Sample: impact on data quality. Statistics Canada, internal document.

Friedman, H.P., and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159-1178.

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.

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.

Gouzi, N., Turmelle, C., Thompson, G. and Rodrigue, J.-F. (2004). Processus de traitement des données du Registre des Adresses. Statistics Canada, internal document.

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

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

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

Kennedy B., Drew J. D., and Lorenz P. (1994). The Impact of Nonresponse Adjustment on Rotation Group Bias in the Canadian Labour Force Survey. Presented at the 5th Atelier international sur la non-réponse dans les enquêtes auprès des ménages, Ottawa, 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.

Labour Force Historical Review. Statistics Canada. Catalogue No. 71-F0004X

Laflamme, G. (2003). Sélection et rotation des UPE : quelques précisions concernant la RAM, Statistics Canada, internal document.

Laflamme, G., and Dochitoiu, C. (2005). An evaluation study for the sequencing of the LFS primary sampling units. Statistics Canada, internal document.

Lebrasseur, D., and Dion, S-M. (2005). The telephone first contact approach in the Labour Force Survey. *Processings of Statistics Canada's Symposium 2005: Methodological Challenges for Future Information Needs.*

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

LFS (2005). Labour Force Survey, annual quality report, January - December 2004. Statistics Canada, Internal document.

LFS-100 CAPI (2006). CAPI Interviewer's Manual. Statistics Canada, internal document, 75030-1800.01

LFS AR Group 1 Clusters Interviewer Manual (2004). CAPI Interviewer's Self-Training Guide, Statistics Canada, internal document, 75030-6489.1

LFS Initial Listing for AR Clusters (2004). CAPI Interviewer's Self-Training Guide, Statistics Canada, internal document, 75030-6458.1

Lindeyer, J. (2004). LFS Mapper User Guide: Inset Changes, Statistics Canada, HSMD, internal document.

Little, R. J. A. (1986). Survey Nonresponse Adjustment for Estimate of Means. *International Statistical Review*, 54, 139-157.

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

Mantel, H., Laniel, N., Duval, M.-C. and Marion, J. (1994). Cost modelling of alternative sample designs for rural areas in the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Mian, I.U.H. and Laniel, J. (1994). Sample allocation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

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

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 117-138.

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.

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

Satin, A., and Shastry, W. (1992). Survey Sampling: A Non-mathematical guide. Second edition, Statistics Canada, catalogue number 12-602-XPE.

Sheridan, M., Drew, D. and Allard, B. (1996). Response rate and the Canadian Labour Force Survey: Luck or Good Planning? *Processings of the Statistics Canada Symposium 96 on Nonsampling Errors*, 65-75.

Simard, M. and Dufour, J. (1995). Impact de l'implantation des interviews assistées par ordinateur comme nouvelle méthode de collecte à l'Enquête sur la population active. Statistics Canada, internal document.

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.

Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey*. Statistics Canada., catalogue no 71-526.

Singh, M.P., Gambino, J. and Mantel, H. (1994). Issues and strategies for small area data. (with discussion). *Survey Methodology*, 20, 3-14.

Sunter, D., Kinack, M., Akyeampong, E. and Charette, D. (1995). Redesigning the Canadian Labour Force Survey Questionnaire: Development and Testing. Statistics Canada, internal document.

Swain, L., Drew, J.D., Lafrance, B. and Lance, K. (1992). The creation of a residential address register for coverage improvement in the 1991 canadian census. *Survey Methodology*, 18, 127-141.

Tambay, J.-L. and Catlin, G. (1995). Sample design of the National Population Health Survey. *Health Reports*, 7, 29-38.

Thompson, G., and Rodrigue, J.-F. (2005). Evaluation of the AR use in the LFS frame. Statistics Canada, internal document.

Thompson, G., and Turmelle, C. (2004). Classification of address register coverage rates - a field study. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Wolter K.M., (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Appendix A.1 Glossary

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

2001 Census Dictionary http://www12.statcan.ca/english/census06/reference/dictionary/index.cfm
Geography illustrated glossary http://geodepot.statcan.ca/Diss2006/Reference/COGG/Index_e.jsp
Guide to the Labour Force Survey http://www.statcan.ca/english/freepub/71-543-GIE/71-543-GIE2007001.htm

Address Register (AR)
A database of residential addresses maintained largely for the Census, but useful to household surveys. The database contains over 11,000,000 addresses, concentrated in major urban areas but covering most communities with civic-style dwelling descriptions.

Address Register (AR) Group
The AR 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.

- 1: Excellent quality, no need to verify the dwelling list, simply select from the current list.
- 2: Good quality, but send the list for verification prior to selecting dwellings.
- 3: Poor quality, or no AR coverage, therefore list from scratch.

Allocation
The process of portioning out a fixed sample size into various provincial and/or sub-provincial areas to satisfy various constraints on collection costs and reliability of estimates

Area frame
A frame of units based on areal extent, such as city blocks, Census Dissemination Areas or similar geography.

Block
A block is an area bounded on all sides by roads and/or boundaries of selected standard geographic areas. Blocks are used to generate larger area units such as Dissemination Areas and LFS PSUs.

Collective dwelling
A collective dwelling is a structure of a commercial, institutional or communal nature where people can reside but where the concept of a single family dwelling is difficult to apply. There are two basic types of collective dwellings: institutional and non-institutional.

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.

Census sub-division
Area that is a municipality or an area that is deemed to be equivalent to a municipality for statistical reporting purposes (*e.g.*, as an Indian reserve or an unorganized territory). Municipal status is defined by laws in effect in each province and territory in Canada.

Dissemination area
Small area composed of one or more neighbouring dissemination blocks, with a population of 400 to 700 persons. All of Canada is divided into dissemination areas. Defined by the Census in much the same manner that LFS PSU were defined by GArDs. Created as a unit for dissemination of Census data.

Design effect
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. These sub-provincial regions are recognized by both the LFS and the Census.

Employment insurance economic region
A set of regions across the country defined by Human Resources Skills Development Canada for the purpose of distributing Employment Insurance benefits in an equitable manner. The LFS is responsible for producing timely estimates required by HRSDC 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:

- 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 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
- had a job but were not at work due to factors such as own illness or disability, personal or family responsibilities, vacation, labour dispute or other reasons (excluding persons on layoff, between casual jobs, and those with 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.

Generalized area delineation system
A set of geography programs used to delineate LFS PSUs in all provinces. Originally created to delineate Enumeration Areas for the Census and then modified by Geography Division for LFS requirements.

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 familiies, 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 whereby the dwellings that belong to one area (PSU, DA) are recorded on paper or electronically. 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 blockfaces 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, or directly from the 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 previous listing efforts. Updates are sent directly to Head Office without involving senior interviewers, unless subsampling is requested due to significant growth.

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.

Primary sampling unit
Units selected at the first stage of sampling in a multistage design are called primary sampling units, or PSUs. The LFS used the Generalized Area Delineation System to form its PSUs.

Randomized probability proportional to size systematic sampling

- In probability proportional to size (PPS) sampling, each sampling unit has a size measure (dwellings in the case of the LFS), and the relative size of units determines their probability of selection. Larger units are more likely to be selected.
- Systematic sampling is a method of constructing 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 (inverse sampling rate).
- The method used by the LFS is called randomized since the list of primary sampling units is randomly ordered prior to sampling systematically.

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?"

Road network
The Road Network is a digital representation of Canada's national road network, containing information such as street names, type, direction and address ranges. Applications include mapping, geo-coding, geographic searching, area delineation, and database maintenance as a source for street names and locations. Since statistical activities do not require absolute positional accuracy, relative positional accuracy takes precedence in the Road Network. As a result, the road network is not suitable for engineering applications, emergency dispatching services, surveying or legal applications.

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 dwellings (or the PSUs that contain them) that rotate in the same month a referred to as a rotation panels. Each panel consists of 1/6th of the sample. As a result, each month has a mix of dwellings in their first, second, third, fourth, fifth and sixth interview. Most strata have 6 (sometimes 12) PSUs selected in order to make each panel as representative of the total as possible. The rotation panel that has its first interviews in a particular month is referred to as the birth panel or birth rotation. Rotation numbers are assigned such that rotation 1 is birthed in January and July, rotation 2 in February and August, up to rotation 6 in June and December. The term off-rotation indicates the introduction of a dwelling with a particular rotation number for a month that does not correspond to the normal pattern for birth rotations.

Rural area
Rural areas include all territory lying outside urban areas. Taken together, urban and rural areas cover all of Canada.

Sampling rate
The ratio of the size of the sample to the size of the population in 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
A measure of variation of a statistic, calculated as the average value of the squared difference of the statistic from its mean over all possible samples.

Slippage rate
A measure of discrepancy between an estimate of population size and the corresponding Census-projected value. It equals 1 – (ratio of estimate to projection).

Stratification
Stratification groups the PSUs created by the GArDs program into conveniently sized sets. The PSUs in a given set (stratum) tend to have similar characteristics. Stratification respects many of the geopolitical boundaries such as province, EIER and Economic Region. Strata are the basic unit in the LFS from which sample selection begins.

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 (inverse sampling rate). With sample rate 1 in 10, and a random starting point of 7, the 7th unit is selected, and every 10th unit thereafter is selected, including 17th, 27th, 37th, *etc.*

Target, population
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
Averages the estimates from the most recent three months, every month.

Two-stage sampling
In two-stage sampling, units are selected in two stages, with the units in the two stages being distinct entities (e.g., blocks in the first stage and dwellings in the second stage). In the first stage of sampling in the LFS, PSUs are selected within strata. In the second stage, dwellings are selected within PSU.

Unemployment
Unemployed persons are those who, during reference week:

- were on temporary layoff during the reference week with an expectation of recall and were available for work, or
- were without work, had actively looked for work in the past four weeks, and were available for work, or
- had a new job to start within four weeks from reference week, and were available for work.

Unemployment rate
The number of unemployed people expressed as a percentage of the active population, or labour force (employed + unemployed). This rate is one the principal statistics produced by the LFS.

Urban area
The Census definition of urban area is an "area with a population of at least 1,000 and no fewer than 400 persons per square kilometre." LFS urban areas are more often delineated by stratum and can be quite different from Census. Urban areas too small for the creation of separate strata were deemed to be rural.

Vacancy rate
The proportion of dwellings that are unoccupied. 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
CD	Census division
CMA	Census metropolitan area
CSD	Census subdivision
CT	Census tract
CV	Coefficient of variation
DA	Dissemination area (census)
EA	Enumeration area (census)
EIER	Employment insurance economic region
ER	Economic Region
FSU	First stage unit (= PSU)
GARDS	Generalized Area Delineation System
HHD	Household
HRSDC	Human Resources Skills Development Canada
ISD	Interviewer Selected Dwellings
ISR	Inverse sampling ratio
LFS	Labour Force Survey
PPS	Probability proportional to size
PSU	Primary sampling unit
RHC	Rao-Hartley-Cochran (random group method)
RO	Regional Office
RPPSS	Randomized PPS systematic
SSU	Secondary sampling unit
UC	Urban centre

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 HRSDC	Total sample
			number		
Newfoundland and Labrador	188,136	907	1,986	18	2,004
Prince Edward Island	50,654	144	1,421	0	1,421
Nova-Scotia	357,712	2,310	2,609	353	2,962
New Brunswick	281,526	2,295	2,604	235	2,840
Quebec	2,970,336	9,466	5,457	4,618	10,075
Ontario	4,226,887	20,165	7,171	8,722	15,893
Manitoba	418,140	14,510	3,254	520	3,774
Saskatchewan	367,815	11,863	3,409	488	3,897
Alberta	1,103,438	10,553	4,030	317	4,347
British Columbia	1,510,709	25,552	4,111	1,329	5,440
Canada	11,475,353	97,765	36,052	16,600	52,653

Table B.2 Number of households in remote strata, high-vacancy strata, and three-stage strata, by province

Province	Remote strata		High-vacancy strata		Three-stage strata	
	Strata	Households	Strata	Households	Strata	Households
			number			
Newfoundland and Labrador	1	3,012	1	5,184	0	0
Prince Edward Island	0	0	0	0	0	0
Nova-Scotia	0	0	1	4,196	0	0
New Brunswick	0	0	1	1,904	0	0
Quebec	1	11,373	2	23,546	3	16,192
Ontario	2	21,595	3	43,991	2	48,575
Manitoba	1	5,310	1	5,119	0	0
Saskatchewan	1	3,988	1	7,077	0	0
Alberta	2	15,673	2	14,213	2	8,529
British Columbia	2	13,390	3	26,253	1	10,231
Canada	10	74,341	15	131,483	8	83,527

Table B.3.1 Statistics for the high-income strata

Census Metropolitan Area (CMA)	High-income strata	Households in high-income strata	Prevalence of high-income households in high-income strata	Prevalence of high-income households in the CMA	High-income households in the CMA that are in to a high-income stratum
	number		percentage		
St. John's	1	4,462	21.4	5.5	26.9
Halifax	2	11,327	23.7	5.8	31.9
Moncton	1	5,116	14.9	4.2	38.5
Saint John	1	5,111	16.4	4.4	39.2
Saguenay	1	4,703	9.8	2.8	26.9
Quebec	1	20,380	21.8	4.1	36.1
Sherbrooke	2	7,327	10.0	2.9	32.8
Trois-Rivières	1	4,055	16.4	2.8	40.3
Montreal	5	118,543	28.1	5.8	40.3
Gatineau	1	6,173	25.3	6.4	23.4
Ottawa	1	18,020	43.3	14.2	17.5
Kingston	1	3,234	22.6	6.2	20.2
Oshawa	1	6,243	33.3	10.5	18.6
Toronto	6	133,184	44.0	13.6	26.3
Hamilton	1	15,728	35.0	9.1	23.8
St. Catharines - Niagara	1	8,224	18.2	5.3	18.9
Kitchener	1	10,029	34.4	8.8	25.2
Brantford	1	4,966	17.1	6.1	31.2
Guelph	1	5,448	23.0	9.0	31.6
London	2	18,428	28.1	6.8	42.5
Windsor	1	8,237	31.5	10.2	21.6
Greater Sudbury	1	3,941	21.0	5.8	22.4
Thunder Bay	1	3,594	17.2	5.2	23.9
Winnipeg	3	19,942	26.6	5.1	38.3
Regina	2	7,830	24.2	5.8	42.5
Saskatoon	1	5,167	21.5	4.8	25.9
Calgary	2	30,814	43.0	12.1	30.5
Edmonton	2	33,571	28.4	7.7	34.6
Abbotsford	1	3,849	14.7	5.2	21.3
Vancouver	4	56,335	30.0	8.8	25.2
Victoria	1	7,113	22.6	5.8	20.5

Table B.3.2 Statistics for the immigrant strata

CMA	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 area in an immigrant stratum
	number		percentage		
Montreal	5	118,828	21.8	4.2	43.4
Ottawa	1	17,868	38.3	7.9	23.2
Toronto	6	132,908	37.9	10.9	28.3
Calgary	2	32,017	24.8	8.5	25.2
Vancouver	4	56,330	31.5	9.5	24.6

1. An immigrant household is a household for which at least one member is an immigrant. For Montreal, Toronto and Vancouver, an immigrant is a person who has immigrated in the last five years. For Ottawa and Calgary, an immigrant is a person who has immigrated in the last ten years.

Table B.3.3 Statistics for the Aboriginal strata

Province	Aboriginal strata	Households in the Aboriginal strata	Aboriginal households[2] in the Aboriginal strata	Aboriginal households in the CMA	Aboriginal households in the CMA that area in an Aboriginal stratum
	number			percentage	
Saskatchewan	6	28,741	38.7	13.1	19.8
Manitoba	7	27,018	34.6	11.8	21.1
Alberta	7	83,137	17.5	6	22.7
British Columbia	11	144,004	14.2	5.2	26.4

2. An Aboriginal household is one in which at least one member is Aboriginal.

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Newfoundland and Labrador	**967**	**37**	**499,608**	**219,337**	**184,842**	**100.0**	**1,944**	**100.0**
EIER								
01	310	14	170,848	68,346	64,133	34.7	664	34.2
02	657	23	328,760	150,991	120,709	65.3	1,280	65.8
ER								
1010	462	20	242,562	102,117	90,001	48.7	932	47.9
1020	89	3	42,133	18,387	15,309	8.3	195	10.0
1030	183	7	100,326	44,592	36,982	20.0	380	19.5
1040	233	7	114,587	54,241	42,550	23.0	437	22.5
Urban Centre[1]								
St. John's (CMA)	314	…	175,918	69,118	64,831	35.1	671	34.5
Corner Brook (UC)	43	…	25,790	10,769	9,833	5.3	103	5.3
Other Urban	69	…	35,819	14,131	13,142	7.1	138	7.1
Non-Urban	541	…	265,081	125,319	97,036	52.5	1,032	53.1
Stratum type								
Regular	909	35	472,835	198,394	175,196	94.8	1,870	96.2
High income	21	1	12,948	4,751	4,462	2.4	45	2.3
High vacancy	37	1	13,825	16,192	5,184	2.8	29	1.5
Prince Edward Island	**286**	**22**	**134,876**	**55,845**	**50,654**	**100.0**	**1,378**	**100.0**
EIER								
03	286	22	134,876	55,845	50,654	100.0	1,378	100.0
ER								
1110	286	22	134,876	55,845	50,654	100.0	1,378	100.0
Urban Centre[1]								
Charlottetown (UC)	116	…	58,211	24,069	22,351	44.1	608	44.1
Summerside (UC)	32	…	16,200	6,593	6,356	12.5	173	12.6
Non-Urban	138	…	60,465	25,183	21,947	43.3	597	43.3
Stratum type								
Regular	286	22	134,876	55,845	50,654	100.0	1,378	100.0
Nova Scotia	**1,681**	**57**	**898,763**	**400,740**	**365,459**	**100.0**	**2,873**	**100.0**
EIER								
04	316	16	180,860	81,940	70,150	19.2	615	21.4
05	713	20	368,980	170,482	147,957	40.5	1,115	38.8
06	652	21	348923	148,318	147,352	40.3	1,143	39.8
ER								
1210	241	10	141,585	63,366	55,199	15.1	486	16.9
1220	306	9	156,394	71,911	61,866	16.9	457	15.9
1230	226	6	120,048	52,514	47,166	12.9	369	12.8
1240	236	7	121,640	59,623	49,771	13.6	385	13.4
1250	672	25	359,096	153,326	151,457	41.4	1,176	40.9

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Nova Scotia (continued)								
Urban Centre[1]								
Halifax (CMA)	672	…	359,096	153,326	151,457	41.4	1,176	40.9
Cape Breton (UC)	175	…	105,968	45,491	41,609	11.4	364	12.7
Truro (UC)	79	…	43,455	19,270	17,434	4.8	123	4.3
New Glasgow (UC)	71	…	36,341	15,826	14,333	3.9	102	3.6
Other Urban	82	…	38,999	18,063	16,831	4.6	130	4.5
Non-Urban	602	…	314,904	148,764	123,795	33.9	978	34.0
Stratum type								
Regular	1,601	54	856,537	378,069	349,936	95.8	2,760	96.1
High income	58	2	32,115	11,910	11,327	3.1	92	3.2
High vacancy	22	1	10,111	10,761	4,196	1.1	21	0.7
New Brunswick	**1,414**	**54**	**721,400**	**310,412**	**282,671**	**100.0**	**2,754**	**100.0**
EIER								
07	666	25	351,009	147,304	139,156	49.2	1,236	44.9
08	251	12	119,028	51,510	462,250	16.4	647	23.5
09	497	17	251,363	111,598	97,265	34.4	871	31.6
ER								
1310	323	13	167,215	70,195	64,284	22.7	578	21.0
1320	353	12	180,906	79,375	71,546	25.3	639	23.2
1330	336	14	167,022	71,455	65,547	23.2	643	23.3
1340	230	7	123,540	54,311	49,387	17.5	452	16.4
1350	172	8	82,717	35,076	31,907	11.3	442	16.0
Urban Centre[1]								
Moncton (UC)	225	…	117,681	48,920	46,801	16.6	420	15.3
Saint John (CMA)	243	…	122,678	51,775	48,262	17.1	433	15.7
Fredericton (UC)	142	…	80,534	34,089	33,210	11.7	286	10.4
Bathurst (UC)	46	…	23,815	10,443	9,550	3.4	86	3.1
Miramichi (UC)	44	…	24,425	10,268	9,392	3.3	84	3.1
Edmundston (UC)	43	…	22,038	9,879	9,014	3.2	125	4.5
Other Urban	45	…	20,922	9,123	8,529	3.0	88	3.2
Non-Urban	626	…	309,307	135,915	117,913	41.7	1,232	44.7
Stratum type								
Regular	1,357	51	686,589	297,914	270,540	95.7	2,645	96.0
High cost	11	1	5,038	2,040	1,904	0.7	17	0.6
High income	46	2	29,773	10,458	10,227	3.6	92	3.3
Quebec	**14,598**	**208**	**7,204,393**	**3,219,007**	**3,033,553**	**100.0**	**9,773**	**5.9**
EIER								
10	290	11	147,382	65,797	58,721	1.9	577	5.9
11	1,415	15	681,517	310,287	303,795	10.0	672	6.9
12	296	14	137,361	64,414	60,012	2.0	651	6.7
13	289	11	145,548	60,578	55,426	1.8	638	6.5
14	313	17	151,605	71,319	65,886	2.2	793	8.1
15	960	14	466,604	203,012	187,623	6.2	747	7.6
16	6,869	56	3,429,391	1,474,891	1,467,550	48.4	2,199	22.5
17	1,996	19	949,056	478,823	392,611	12.9	924	9.5
18	479	8	233,747	114,748	93,577	3.1	498	5.1
19	888	12	448,950	200,062	178,335	5.9	727	7.4
20	498	17	258,294	109,958	107,820	3.6	758	7.8
21	305	14	154,938	65,118	62,197	2.1	589	6.0

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Quebec (continued)								
ER								
2410	178	7	94,994	42,050	37,643	1.2	409	4.2
2415	430	7	200,630	94,444	81,830	2.7	403	4.1
2420	1,350	18	637,677	299,210	286,618	9.4	637	6.5
2425	772	16	383,376	161,768	150,565	5.0	817	8.4
2430	600	19	285,613	136,776	120,369	4.0	1,006	10.3
2433	442	7	217,931	94,488	87,619	2.9	249	2.5
2435	2,474	33	1,277,949	527,886	507,854	16.7	1,219	12.5
2440	3,929	26	1,814,117	844,746	841,285	27.7	1,153	11.8
2445	633	5	343,419	135,811	139,591	4.6	271	2.8
2450	725	5	386,824	167,892	146,756	4.8	286	2.9
2455	884	8	460,952	218,799	181,720	6.0	369	3.8
2460	634	17	315,181	147,139	131,340	4.3	862	8.8
2465	286	7	142,329	63,462	57,788	1.9	329	3.4
2470	545	14	252,338	120,600	110,115	3.6	753	7.7
2475	536	15	275,845	114,451	108,207	3.6	754	7.7
2480	145	3	89,460	40,178	35,773	1.2	206	2.1
2490	35	1	25,758	9,307	8,480	0.3	50	0.5
Urban Centre[1]								
Chicoutimi (CMA)	305	…	154,938	65,118	62,197	2.1	589	6.0
Quebec (CMA)	1,415	…	681,517	310,287	303,795	10.0	672	6.9
Sherbrooke (CMA)	317	…	153,811	72,218	66,731	2.2	803	8.2
Trois-Rivieres (CMA)	296	…	137,361	64,414	60,012	2.0	651	6.7
Shawinigan (UC)	122	…	57,304	28,543	25,601	0.8	54	0.6
Drummondville (UC)	140	…	68,451	30,266	28,948	1.0	70	0.7
Granby (UC)	130	…	60,264	26,585	25,285	0.8	101	1.0
Saint-Jean-sur-Richelieu (UC)	164	…	79,919	34,014	32,797	1.1	129	1.3
Montreal (CMA)	6,869	…	3,429,391	1,474,891	1,467,550	48.4	2,199	22.5
Gatineau (CMA)	498	…	258,294	109,958	107,820	3.6	758	7.8
Rimouski (UC)	91	…	47,688	22,277	20,040	0.7	79	0.8
Baie-Comeau (UC)	57	…	28,940	12,831	11,615	0.4	72	0.7
Sept-Iles (UC)	43	…	24,721	11,313	10,204	0.3	63	0.6
Rouyn-Noranda/Val-d'Or (UC)	130	…	67,657	30,835	28,165	0.9	158	1.6
Other Urban	721	…	396,528	177,781	166,701	5.5	761	7.8
Non-Urban	3,300	…	1,557,609	747,676	616,092	20.3	2,614	26.7
Stratum type								
Regular	13,070	189	6,394,430	2,847,917	2,703,149	89.1	9,082	92.9
High income	779	11	458,544	167,589	163,747	5.4	441	4.5
High vacancy	131	2	51,669	63,903	23,354	0.8	51	0.5
Immigrant	543	5	267,034	126,205	131,930	4.3	169	1.7
Remote	75	1	32,716	13,393	11,373	0.4	30	0.3

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Ontario	**19,905**	**332**	**1,1432,966**	**4,537,466**	**4,403,105**	**100.0**	**15,416**	**100.0**
EIER								
22	1,490	23	822,608	326,805	329,173	7.5	872	5.7
23	779	10	401,080	180,391	157,812	3.6	841	5.5
24	276	12	145,610	63,746	59,295	1.3	677	4.4
25	1,932	23	1,028,112	502,899	395,046	9.0	1,259	8.2
26	504	14	296,770	106,576	108,614	2.5	653	4.2
27	7,631	74	4,747,842	1,693,381	1,785,752	40.6	2,640	17.1
28	1,238	16	664,593	261,718	261,700	5.9	758	4.9
29	744	17	377,009	159,032	152,795	3.5	847	5.5
30	791	19	415,990	177,522	176,612	4.0	872	5.7
31	437	15	254,636	98,157	93,642	2.1	724	4.7
32	568	18	300,501	121,006	116,531	2.6	648	4.2
33	717	16	415,997	159,339	155,664	3.5	702	4.6
34	567	10	314,297	129,721	121,150	2.8	719	4.7
35	908	22	501,012	192,682	186,326	4.2	1,056	6.9
36	291	16	150,461	66,588	62,478	1.4	701	4.5
37	250	13	120,353	53,405	49,001	1.1	608	3.9
38	782	14	476,095	244,498	191,514	4.3	839	5.4
ER								
3510	2,072	30	1,122,635	456,715	447,468	10.2	1,502	9.7
3515	831	19	422,350	195,847	168,989	3.8	1,040	6.7
3520	634	8	337,306	189,030	132,249	3.0	433	2.8
3530	8,058	86	4,996,356	1,784,224	1,878,538	42.7	3,205	20.8
3540	1,847	36	1,062,025	422,275	391,419	8.9	1,660	10.8
3550	2,380	48	1,273,928	510,986	500,710	11.4	2,282	14.8
3560	1,105	27	586,661	242,328	238,877	5.4	1,204	7.8
3570	1,116	24	605,638	247,053	234,257	5.3	1,347	8.7
3580	539	11	279,158	124,517	107,605	2.4	595	3.9
3590	975	27	536,612	268,659	218,931	5.0	1,371	8.9
3595	348	16	210,297	95,832	84,062	1.9	777	5.0
Urban Centre[1]								
Corwall (UC)	101	…	57,581	24,634	23,747	0.5	125	0.8
Ottawa (CMA)_	1,471	…	810,462	322,544	325,020	7.4	849	5.5
Kingston (CMA)	276	…	145,610	63,746	59,295	1.3	677	4.4
Belleville (UC)	182	…	87,395	37,185	35,814	0.8	78	0.5
Peterborough (UC)	198	…	100,738	43,893	39,985	0.9	208	1.3
Kawartha Lakes (UC)	113	…	69,179	34,637	27,182	0.6	64	0.4
Oshawa (CMA)	504	…	296,770	106,576	108,614	2.5	653	4.2
Toronto (CMA)	7,631	…	4,747,842	1,693,381	1,785,752	40.6	2,640	17.1
Hamilton (CMA)	1,238	…	664,593	261,718	261,700	5.9	758	4.9
St-Catharines/Niagara (CMA)	744	…	377,009	159,032	152,795	3.5	847	5.5
Kitchener (CMA)	717	…	415,997	159,339	155,664	3.5	702	4.6
Brantford (UC)	166	…	86,417	34,881	33,849	0.8	315	2.0
Norfolk (UC)	101	…	59,947	24,075	22,526	0.5	141	0.9
Guelph (UC)	216	…	117,345	46,253	47,128	1.1	268	1.7
London (CMA)	828	…	435,104	184,706	183,600	4.2	909	5.9
Chatham-Kent (UC)	178	…	106,864	44,529	42,202	1.0	251	1.6
Windsor (CMA)	582	…	308,735	124,402	119,397	2.7	662	4.3
Sarnia (UC)	173	…	87,636	37,581	35,505	0.8	213	1.4
Barrie (UC)	273	…	156,457	59,227	54,996	1.2	231	1.5
North Bay (UC)	123	…	62,303	27,169	24,896	0.6	111	0.7
Greater Sudbury (CMA)	300	…	155,219	68,690	64,351	1.5	723	4.7
Sault Ste-Marie (UC)	160	…	78,049	34,539	31,919	0.7	140	0.9
Thunder Bay (UC)	252	…	121,387	53,870	49,337	1.1	612	4.0
Other Urban	944	…	607,718	258,188	243,645	5.5	1,184	7.7
Non-Urban	2,434	…	1,276,609	632,671	474,186	10.8	2,055	13.3

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Ontario (continued)								
Stratum type								
Regular	17,739	301	10,086,861	3,964,127	3,923,649	89.1	14,218	92.2
High income	1,174	19	742,037	244,895	242,493	5.5	781	5.1
High vacancy	231	3	102,463	141,673	43,548	1.0	109	0.7
Immigrant	617	7	445,725	154,294	171,820	3.9	257	1.7
Remote	144	2	55,880	32,477	21,595	0.5	51	0.3
Manitoba	**2,120**	71	1,062,777	452,019	431,986	100.0	3,661	100.0
EIER								
30	1,318	41	671,216	280,231	283,789	65.7	1,917	52.4
40	622	20	310,898	130,561	117,015	27.1	1,017	27.8
41	180	10	80,663	41,227	31,182	7.2	727	19.9
ER								
4610	158	9	83,427	31,904	29,044	6.7	359	9.8
4620	98	3	51,703	19,248	18,201	4.2	189	5.2
4630	218	6	100,658	46,576	40,770	9.4	314	8.6
4640	84	4	44,093	16,490	15,236	3.5	156	4.3
4650	1,233	38	621,765	262,002	267,337	61.9	1,800	49.2
4660	148	4	79,463	37,334	29,031	6.7	356	9.7
4670	94	3	41,685	20,794	17,672	4.1	163	4.5
4680	87	4	39,983	17,671	14,695	3.4	324	8.9
Urban Centre[1]								
Winnipeg (CMA)	1,318	…	671,216	280,231	283,789	65.7	1,917	52.4
Brandon (UC)	82	…	41,037	17,966	17,185	4.0	134	3.7
Other Urban	87	…	41,737	18,050	16,985	3.9	309	8.4
Non-Urban	633	…	308,787	135,772	114,027	26.4	1,301	35.5
Stratum type								
Regular	1,806	60	902,291	902,291	371,650	86.0	3,157	86.2
Aboriginal	146	6	73,210	73,210	28,741	6.7	289	7.9
High income	99	3	60,247	60,247	21,166	4.9	143	3.9
High vacancy	28	1	12,097	12,097	5,119	1.2	30	0.8
Remote	41	1	14,932	14,932	5,310	1.2	42	1.1
Saskatchewan	**2,006**	**72**	**928,348**	**415,768**	**369,964**	**100.0**	**3,780**	**100.0**
EIER								
42	397	19	192,800	80,772	77,061	20.8	902	23.9
43	429	17	225,772	94,639	91,053	24.6	799	21.1
44	767	22	314,420	148,951	127,625	34.5	1,289	34.1
45	413	14	195,356	91,406	74,225	20.1	790	20.9
ER								
4710	580	25	266,594	116,521	106,809	28.9	1,183	31.3
4720	239	8	104,048	48,261	41,789	11.3	472	12.5
4730	571	26	282,498	119,785	113,074	30.6	1,011	26.7
4740	211	5	83,672	41,261	35,488	9.6	339	9.0
4750	377	7	179,105	84,786	68,816	18.6	744	19.7
4760	28	1	12,431	5,154	3,988	1.1	31	0.8
Urban Centre[1]								
Regina (CMA)	397	…	192,800	80,772	77,061	20.8	902	23.9
Saskatoon (CMA)	429	…	225,772	94,639	91,053	24.6	799	21.1
Moose Jaw (UC)	64	…	33,616	15,056	14,061	3.8	161	4.3
Prince Albert (UC)	69	…	41,200	15,977	15,133	4.1	168	4.4
Other Urban	124	…	61,248	27,114	25,611	6.9	273	7.2
Non-Urban	923	…	373,712	182,210	147,045	39.7	1,477	39.1
Stratum type								
Regular	1,731	61	790,087	344,936	318,695	86.1	3,296	87.2
Aboriginal	139	6	68,922	29,379	27,018	7.3	283	7.5
High income	67	3	40,105	13,377	13,405	3.6	139	3.7
High vacancy	41	1	16,803	22,922	6,858	1.9	31	0.8
Remote	28	1	12,431	5,154	3,988	1.1	31	0.8

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
Alberta	**5,317**	**85**	**2,948,116**	**1,165,969**	**1,155,109**	**100.0**	**5,416**	**100.0**
EIER								
46	1,699	22	971,117	376,446	388,273	33.6	1,206	22.3
47	1,711	22	938,129	372,728	378,920	32.8	1,392	25.7
48	339	14	198,600	81,163	69,382	6.0	1,010	18.6
49	1,568	27	840,270	335,632	318,534	27.6	1,808	33.4
ER								
4810	447	8	233,502	91,700	88,312	7.6	303	5.6
4820	371	9	179,432	72,365	66,546	5.8	439	8.1
4830	1,824	23	1,040,645	402,812	414,375	35.9	1,291	23.8
4840	116	4	71,928	31,069	27,314	2.4	268	4.9
4850	281	5	151,286	60,696	61,543	5.3	333	6.1
4860	1,776	21	974,496	388,697	392,588	34.0	1,476	27.3
4870	363	10	208,643	83,851	73,839	6.4	766	14.1
4880	139	5	88,184	34,779	30,592	2.6	540	10.0
Urban Centre[1]								
Medicine Hat (UC)	122	…	61,735	25,693	24,372	2.1	77	1.4
Lethbridge (UC)	138	…	67,375	29,276	29,557	2.6	97	1.8
Calgary (CMA)	1,699	…	971,117	376,446	388,273	33.6	1,206	22.3
Red Deer (UC)	135	…	69,608	27,871	32,097	2.8	183	3.4
Edmonton (CMA)	1,711	…	938,129	372,728	378,920	32.8	1,392	25.7
Other Urban	194	…	128,397	51,069	48,189	4.2	397	7.3
Non-Urban	1,318	…	711,755	282,886	253,701	22.0	2,064	38.1
Stratum type								
Regular	4,298	69	2,363,145	942,186	940,942	81.5	3,852	71.1
Aboriginal	422	7	214,115	87,886	83,934	7.3	1,173	21.7
High cost	64	1	27,485	10,935	9,661	0.8	63	1.2
High income	295	4	200,078	66,279	67,105	5.8	180	3.3
High vacancy	26	1	11,089	11,718	4,552	0.4	20	0.4
Immigrant	145	2	94,642	33,372	36,915	3.2	88	1.6
Remote	67	1	37,562	13,593	12,000	1.0	40	0.7
British Columbia	7,311	124	3,838,992	1,615,758	1,558,403	100.0	6,377	100.0
EIER								
50	1,192	13	591,184	270,509	247,068	15.9	954	15.0
51	242	11	147,895	52,706	53,668	3.4	573	9.0
52	3,676	54	1,982,716	784,745	797,388	51.2	2,346	36.8
53	644	18	307,340	139,992	136,267	8.7	914	14.3
54	985	14	490,650	231,587	203,222	13.0	834	13.1
55	572	14	319,207	136,219	120,790	7.8	756	11.9
ER								
5910	1,384	32	671,630	310,359	289,090	18.6	1,542	24.2
5920	4,181	68	2,269,754	904,810	906,594	58.2	3,154	49.5
5930	886	9	447,080	199,614	186,651	12.0	697	10.9
5940	306	3	144,104	70,895	60,417	3.9	260	4.1
5950	292	6	157,742	67,890	60,320	3.9	327	5.1
5960	79	2	51,954	21,849	19,490	1.3	149	2.3
5970	75	1	38,726	16,448	14,306	0.9	89	1.4
5980	108	3	58,002	23,893	21,535	1.4	166	2.6

1. See note at end of table.
Note: See Appendix A.2 for abbreviations

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

Province	PSUs	Strata	Population	Dwellings	Households		Sampled households	
			number			%	number	%
British Columbia (continued)								
Urban Centre[1]								
Kelowna (UC)	270	…	139,882	58,875	58,132	3.7	212	3.3
Vernon (UC)	96	…	49,727	21,583	20,868	1.3	61	1.0
Kamloops (UC)	160	…	84,721	36,040	33,770	2.2	147	2.3
Chilliwack (UC)	120	…	67,359	27,360	26,049	1.7	109	1.7
Abbotsford (CMA)	242	…	147,895	52,706	53,668	3.4	573	9.0
Vancouver (CMA)	3,676	…	1,982,716	784,745	797,388	51.2	2,346	36.8
Victoria (CMA)	644	…	307,340	139,992	136,267	8.7	914	14.3
Nanaimo (UC)	179	…	85,008	37,349	35,395	2.3	147	2.3
Prince George (UC)	159	…	84,929	34,400	32,085	2.1	190	3.0
Dawson Creek (UC)	37	…	17,251	7,154	6,662	0.4	53	0.8
Other Urban	1,305	…	629,094	308,427	257,064	16.5	469	7.4
Non-Urban	423	…	243,070	107,127	101,055	6.5	1,156	18.1
Stratum type								
Regular	5,745	98	3,030,838	1,267,066	1,239,216	79.5	4,382	68.7
Aboriginal	733	11	346,444	153,511	147,029	9.4	1,505	23.6
High cost	78	2	33,910	17,398	13,741	0.9	54	0.8
High income	330	6	197,680	69,909	68,487	4.4	228	3.6
High vacancy	63	1	28,365	33,282	12,383	0.8	28	0.4
Immigrant	274	4	165,880	58,754	64,157	4.1	142	2.2
Remote	88	2	35,875	15,838	13,390	0.9	38	0.6

1. CMA boundaries follow preliminary 2006 Census Metropolitan Areas (as of 2003). Other Urban Centres (UC) follow 2001 Census Agglomerations where available. Urban Strata do not follow Census Urban areas.
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 2001 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.

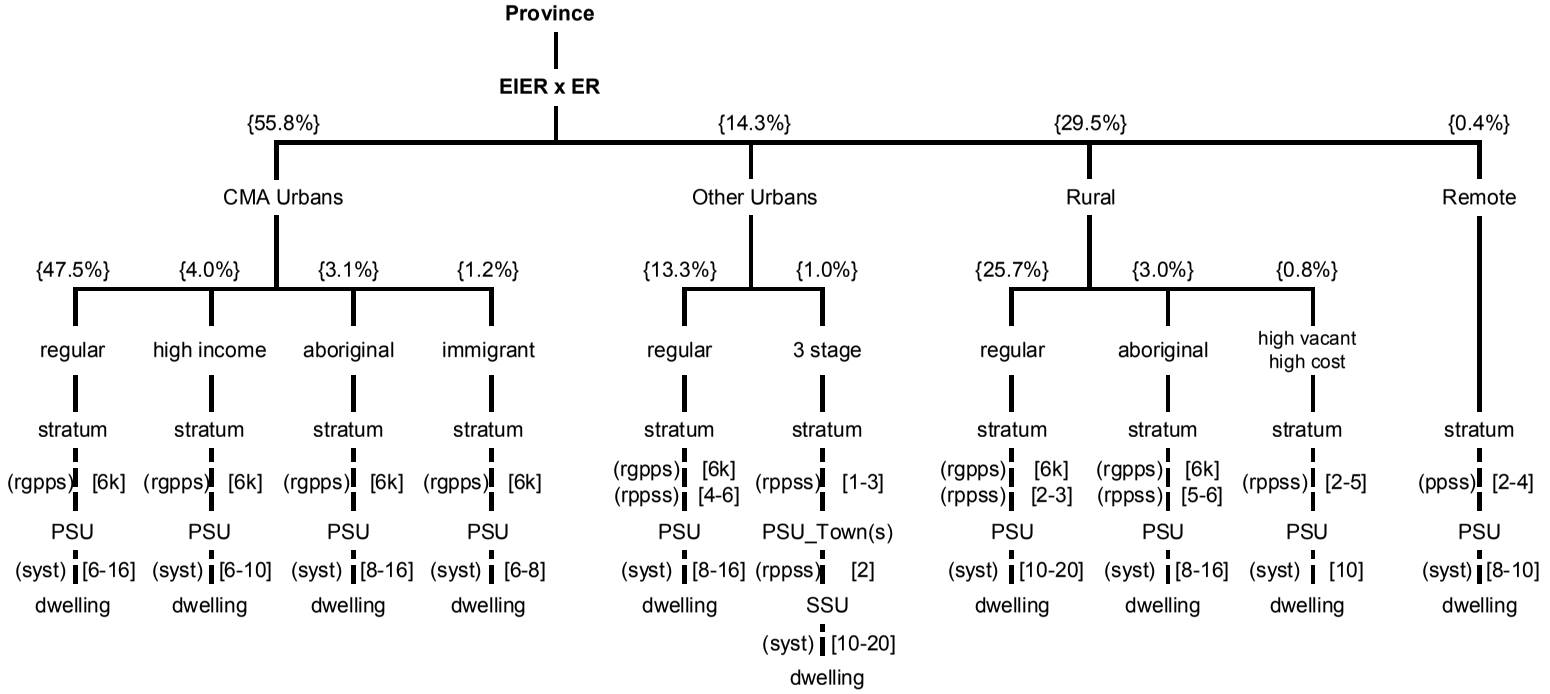
Number of persons employed in the following sectors:

- 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

Labour Force Survey Sample Design - since 2005

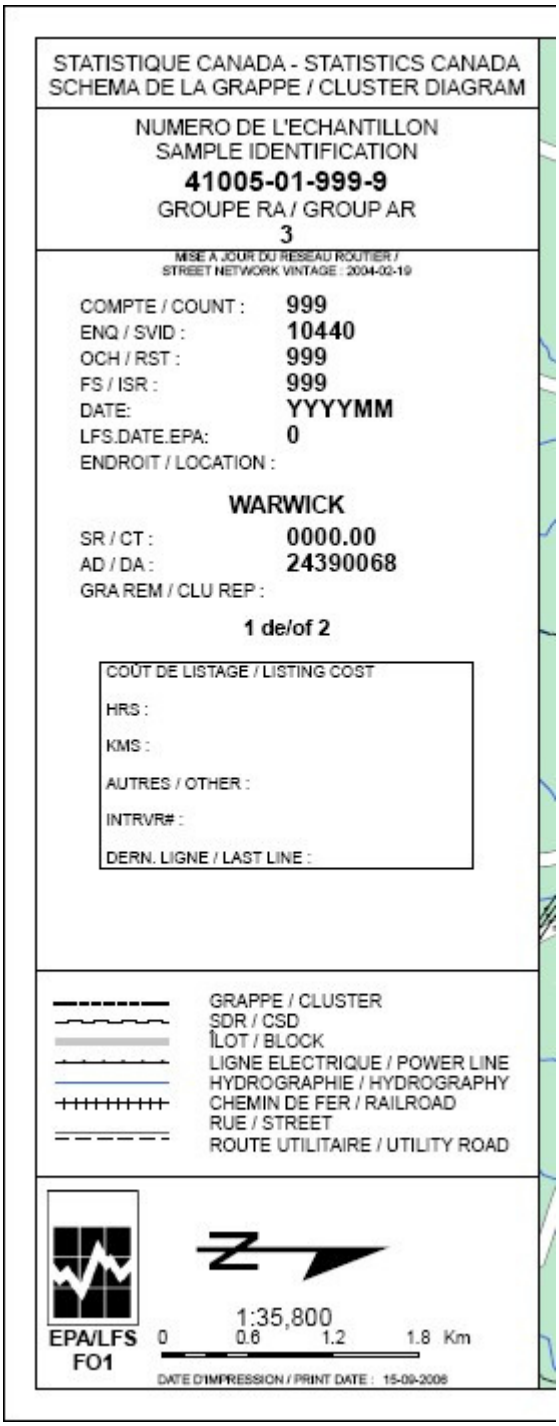


= level of stratification	EIER	Employment Insurance Economic Region	{%} percent of total sample (with top-up and 3% reduction)
	ER	Economic Region	( ) selection method
	CMA	Census Metropolitan Area	[ ] number of units selected (6k = multiple of 6)
= stage of sampling	PSU	Primary Sampling Unit	pps probability proportional to size
	SSU	Secondary Sampling Unit in 3 stage	rgpps random group pps systematic
	cluster	Set of Census Blocks	rppss randomized pps systematic
			ppss pps systematic (ordered list)
			syst systematic

Some regular and aboriginal strata had too few clusters to warrant rgpps.

Appendix D Cluster map examples (form F01)

Before presenting some examples of cluster maps, we explain the F01 legend, from top to bottom.



PSU identification (stratum 41005; design type 01; PSU 999; rotation group 9)

AR Group (3)
Date of latest street network (2004-02-19)
Expected number of households in PSU (999)
First Survey to use the PSU (10440)
Random Start and Inverse Sample Rate to use in dwelling selection (999, 999)
Date of first use (YYYYMM)
Date of first LFS use (usually blank) (0)

General location based on Postal Code Conversion File (Warwick)
2001 Census Tract ID for locating PSU (0000.00)
Predominant 2001 Dissemination Area (24390068)

Map number and number of maps for this PSU (1 of 2)

Blank form for interviewers to record listing costs.

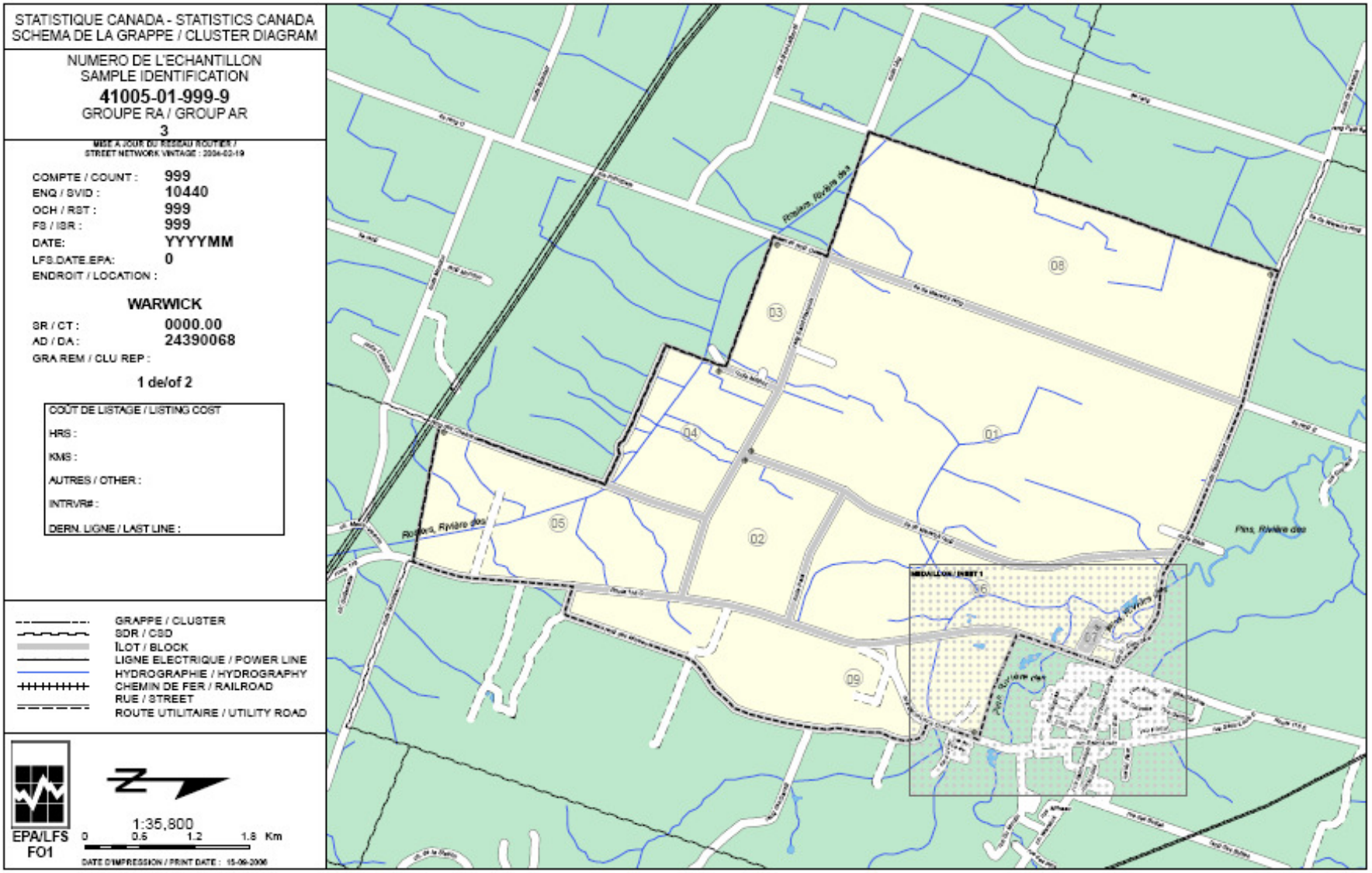
Legend of major geography features:

Blocks are normally outlined with one starting point (circled x) in each.
Area outside cluster is green, area inside is yellow.
Only utility roads (trails) are distinguished separately. Proposed roads are not easily identified or maintained.

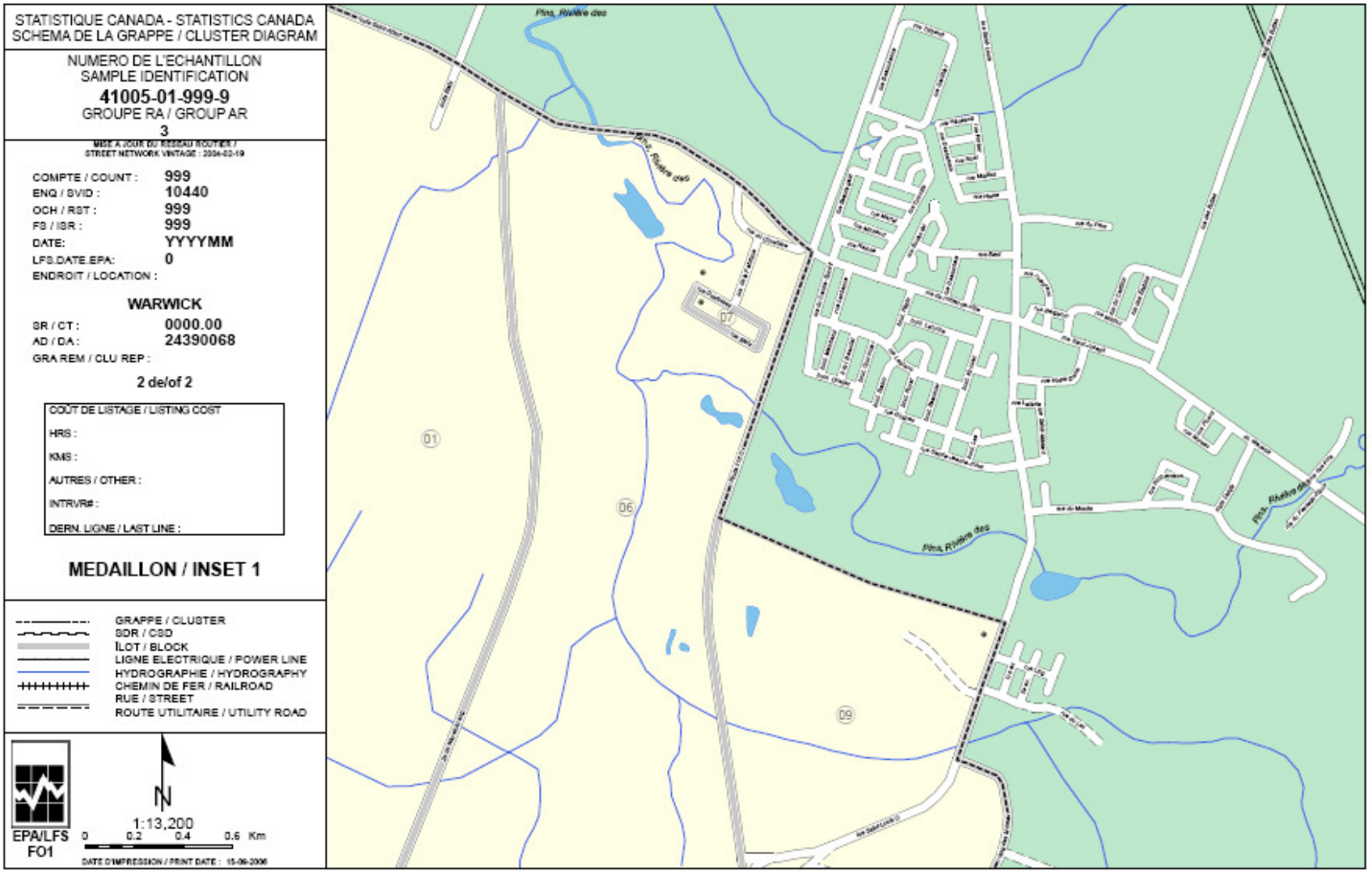
North symbol always points up or to the right for ease in working while map is in interviewer's binder.

Scale is useful to measure driving distance when geographic features are missing at boundary.
Date printed (15-09-2006)

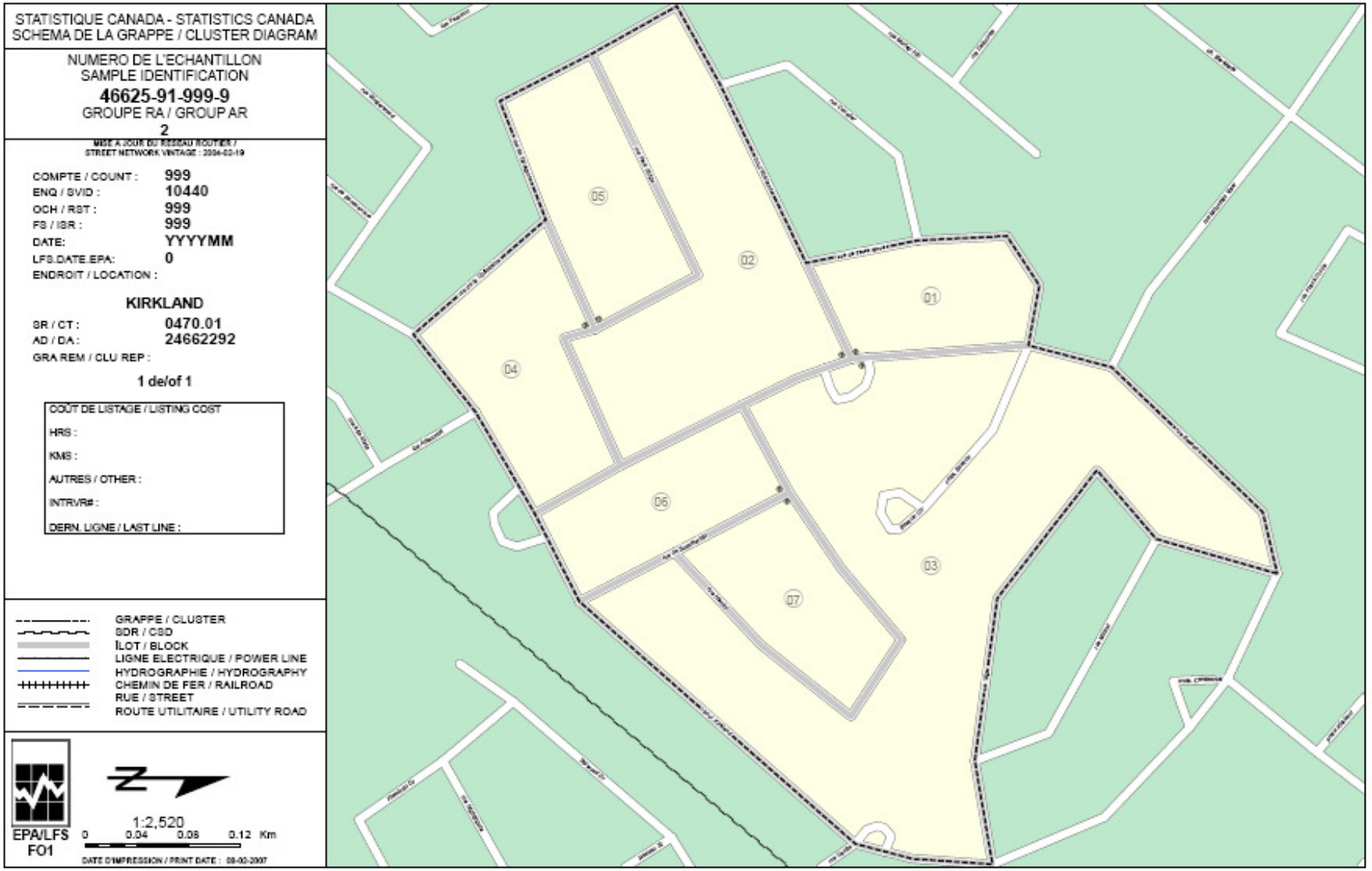
Example 1: Map size 1 (11”x17”), AR Group 3, One Inset. North to the right. There are a few imaginary lines along the top corresponding to the 2001 Census Sub-division limits. The details near the urban area to the north-east requires clarification, hence the inset. Roads that do not form a block will not have the block boundary (thick gray line).



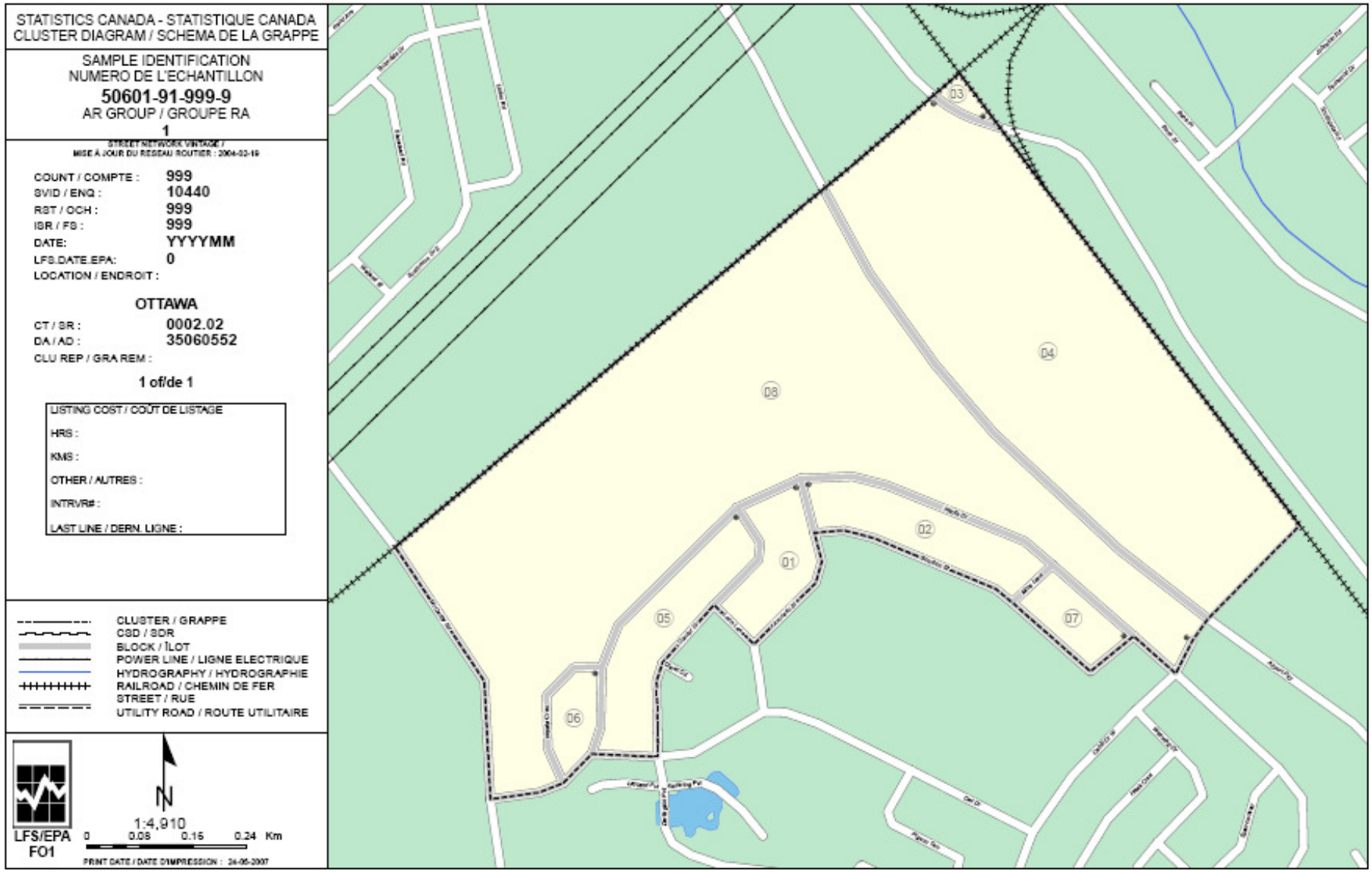
Example 1a: Inset for Example 1, showing that almost all of the small village is outside the area to list. Note the orientation now has North to the top. Inset 1 is map 2 of 2 in this series. The rest of the legend is identical with the main map. Utility roads are not normally a block boundary.



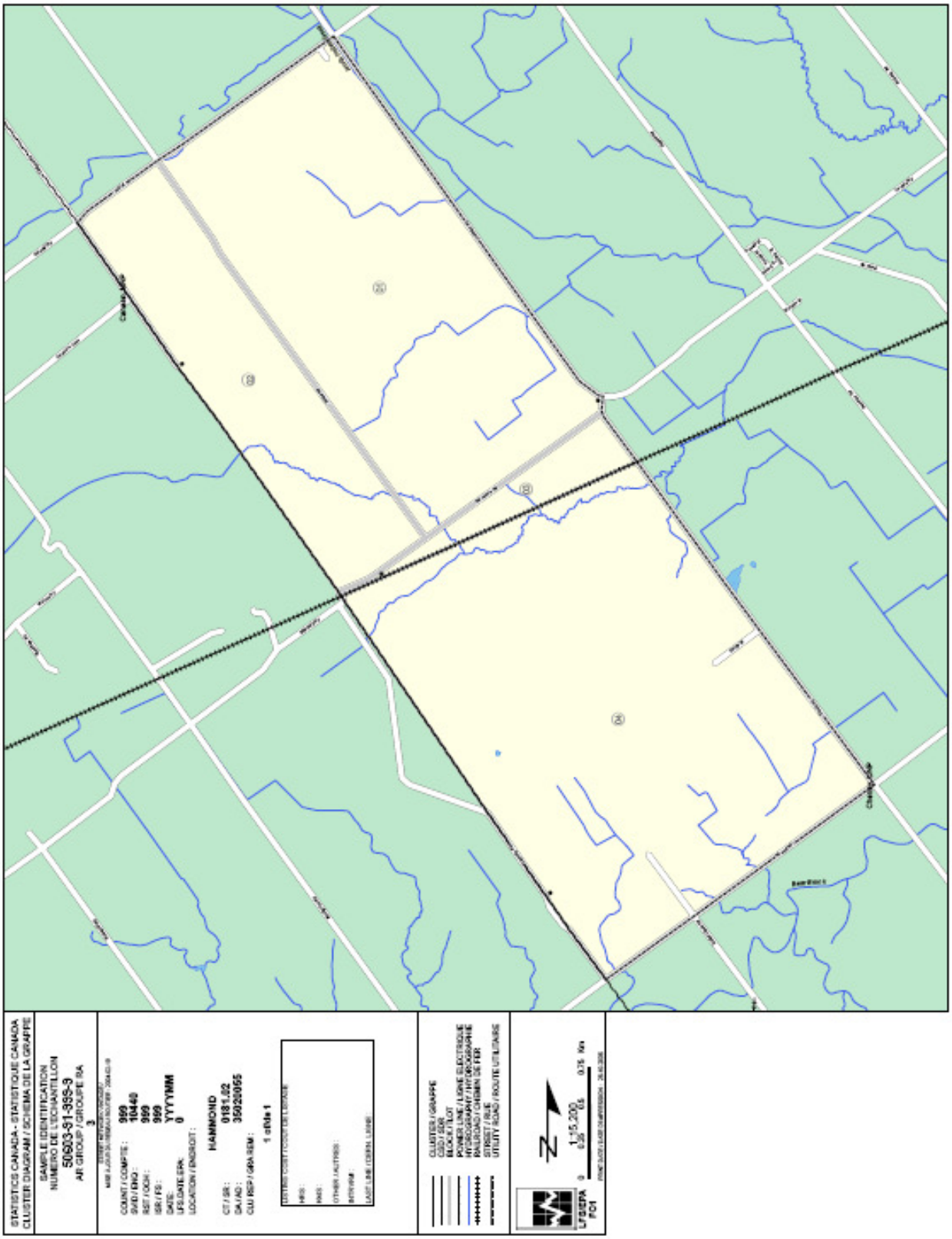
Example 2: Map size 1, AR Group 2, no insets, North to the right. Note that starting points are clustered to reduce travel distance. Not every block has a block boundary (thick gray line). Either the block is too small to have any dwellings, or forms part of a non-addressable area such as highway clover-leafs or indicates new roads since the block formation exercise.



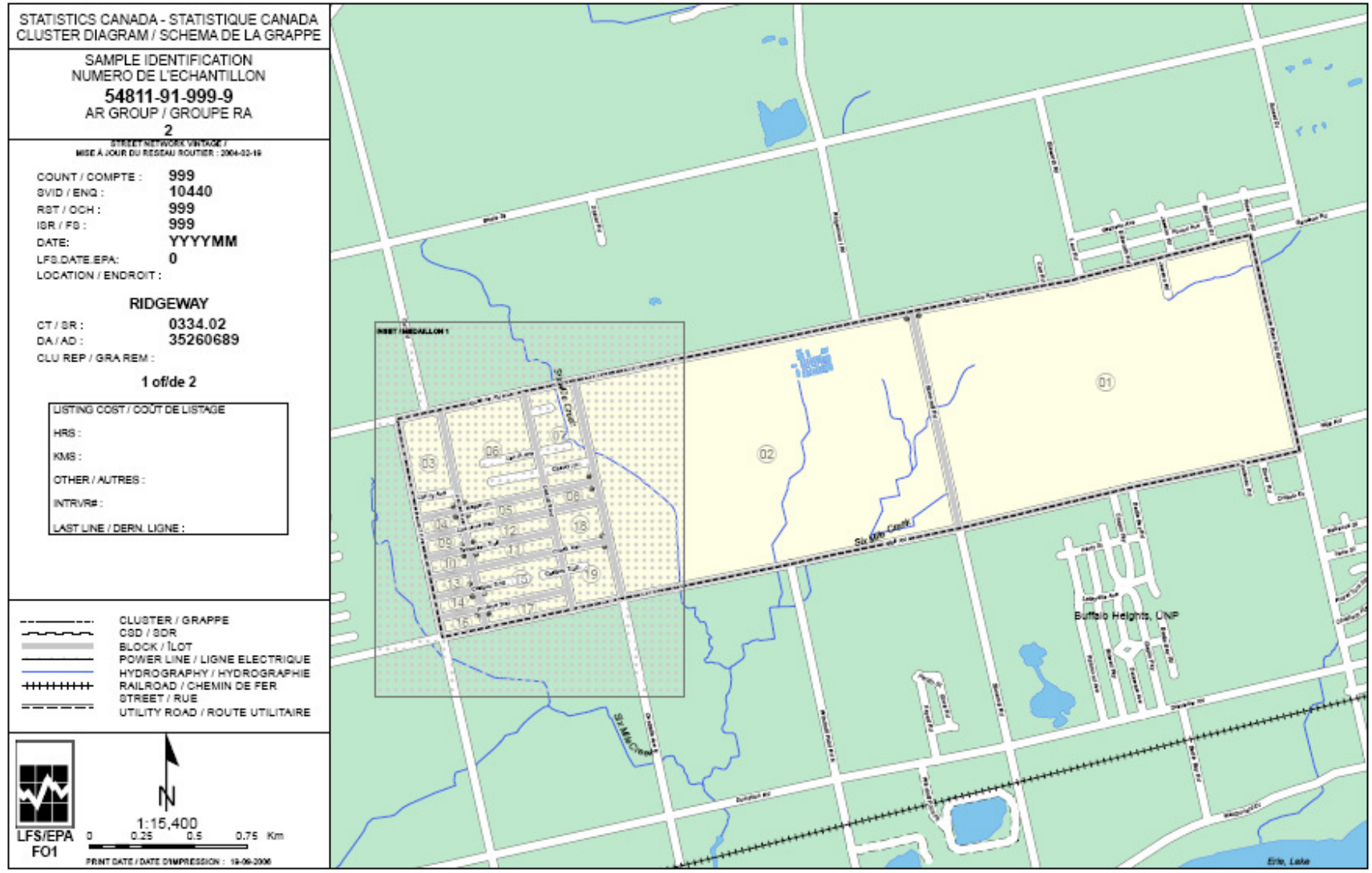
Example 3: Map size 1, AR Group 1, no insets, North to the top. It is not obvious from the location or the road network that this could or should be an AR group 1. Railways can be used as boundaries but generate short segments of roads to include in listing.



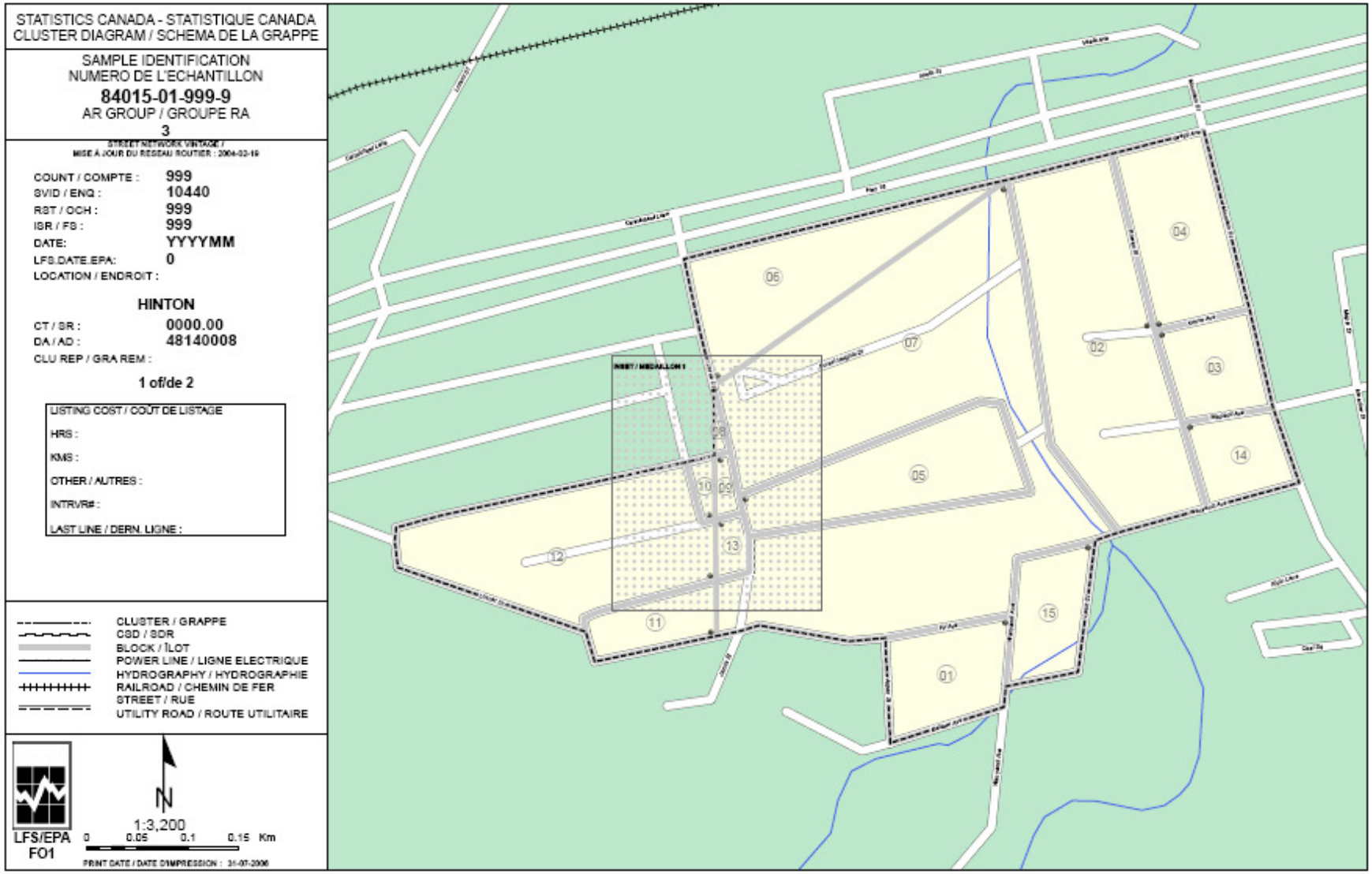
Example 4: Map size 2 (17”x22”), AR Group 3, no insets, North to the right. The larger map sizes are more difficult to display on these pages. Map size 3 (17”x34”) examples are not included. This rural area is probably not covered by the Address Register despite the fact it resides within the CMA of Ottawa. Note the unincorporated place names to aid in locating the area to list.



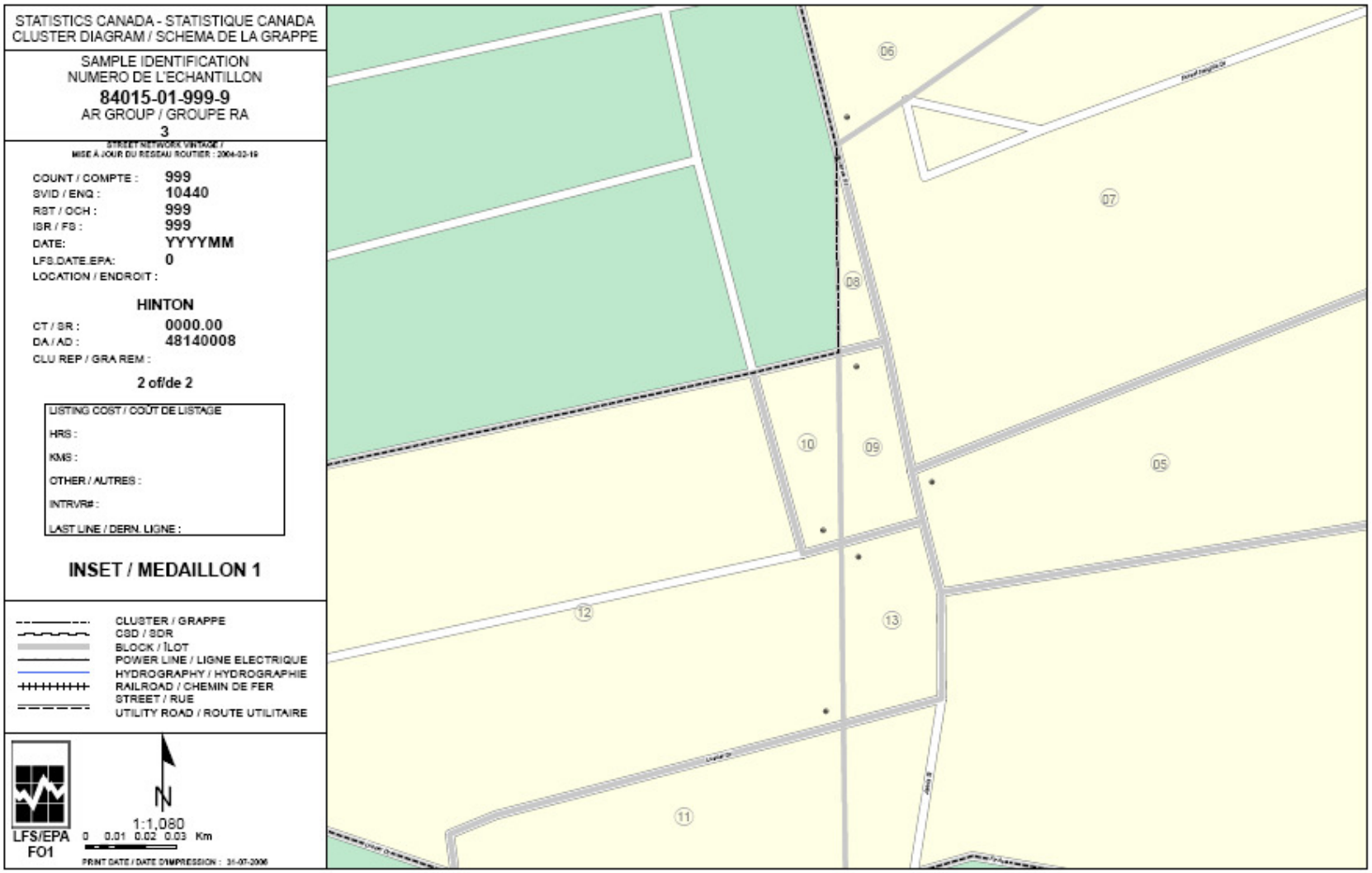
Example 5: Map size 1, AR Group 2, One Inset, North to the top. Many small blocks require an inset. Map sizes and insets are created automatically, but some situations need correction. A map size 2 with no inset may suffice for this map. Inset map not included in examples.



Example 6: Map size 1, AR Group 3, One Inset, North to the top. Imaginary line generated multiple block splits that require an inset. Other roads added too late to be included in block formation. The imaginary line can be a former CSD boundary, FED boundary or even an older EA boundary. New blocks are formed only occasionally. In more complex examples, the block boundaries, numbering and starting points are removed in order to ease the task of listing, especially for AR Group 3 that have no AR pre-listed dwellings.

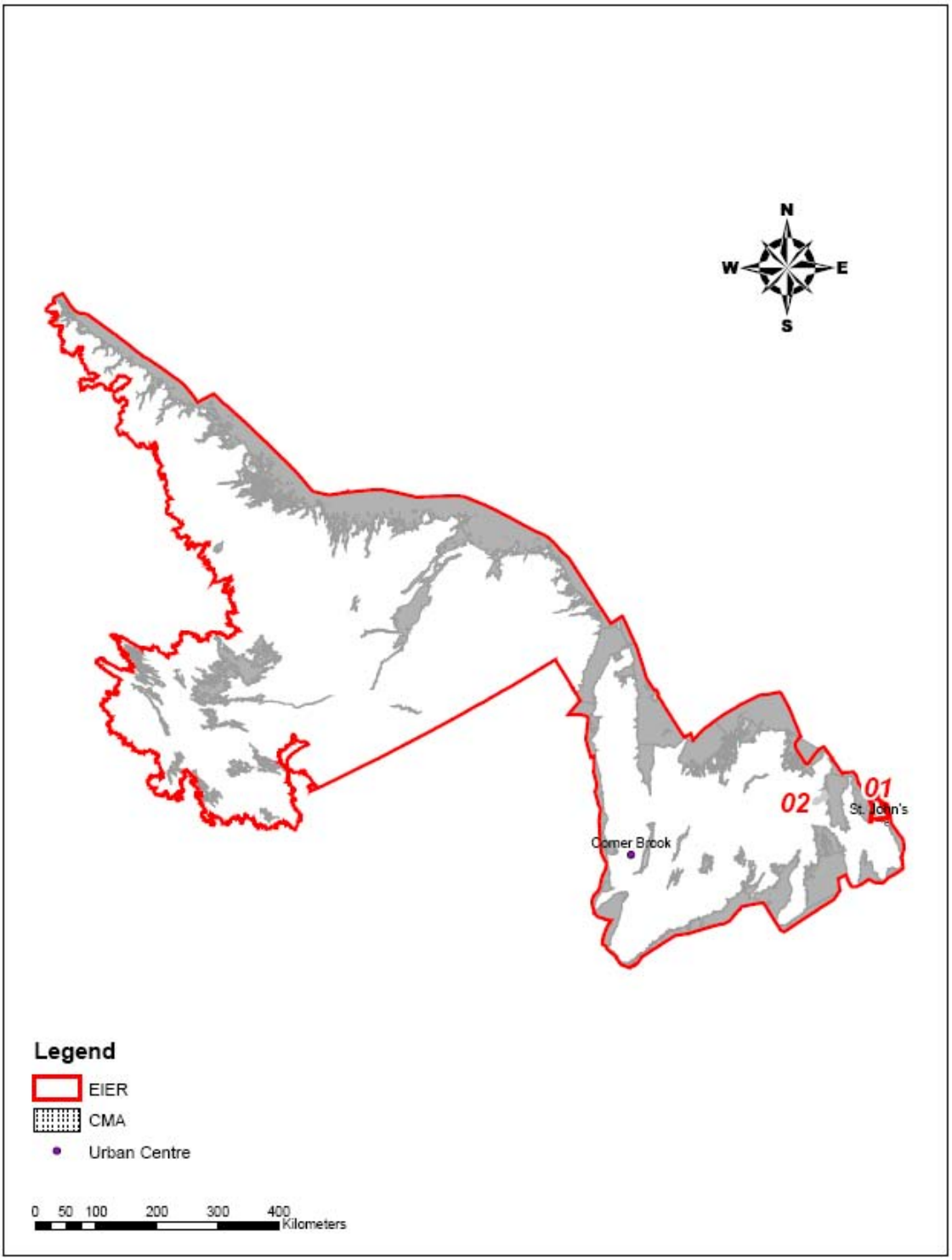


Example 6a: Inset 1 for Example 6. Redefining block boundaries to the latest road network requires re-sequencing blocks and AR dwellings (if available) such as in the merge of blocks 09 and 10. Block 08 would disappear as a separate entity and thus redefines the cluster boundary and the listing required. In this case the addresses in block 08 would likely move to the westerly neighbouring area. Any such dwellings in sample pose an overlap issue if the neighbour is also selected.

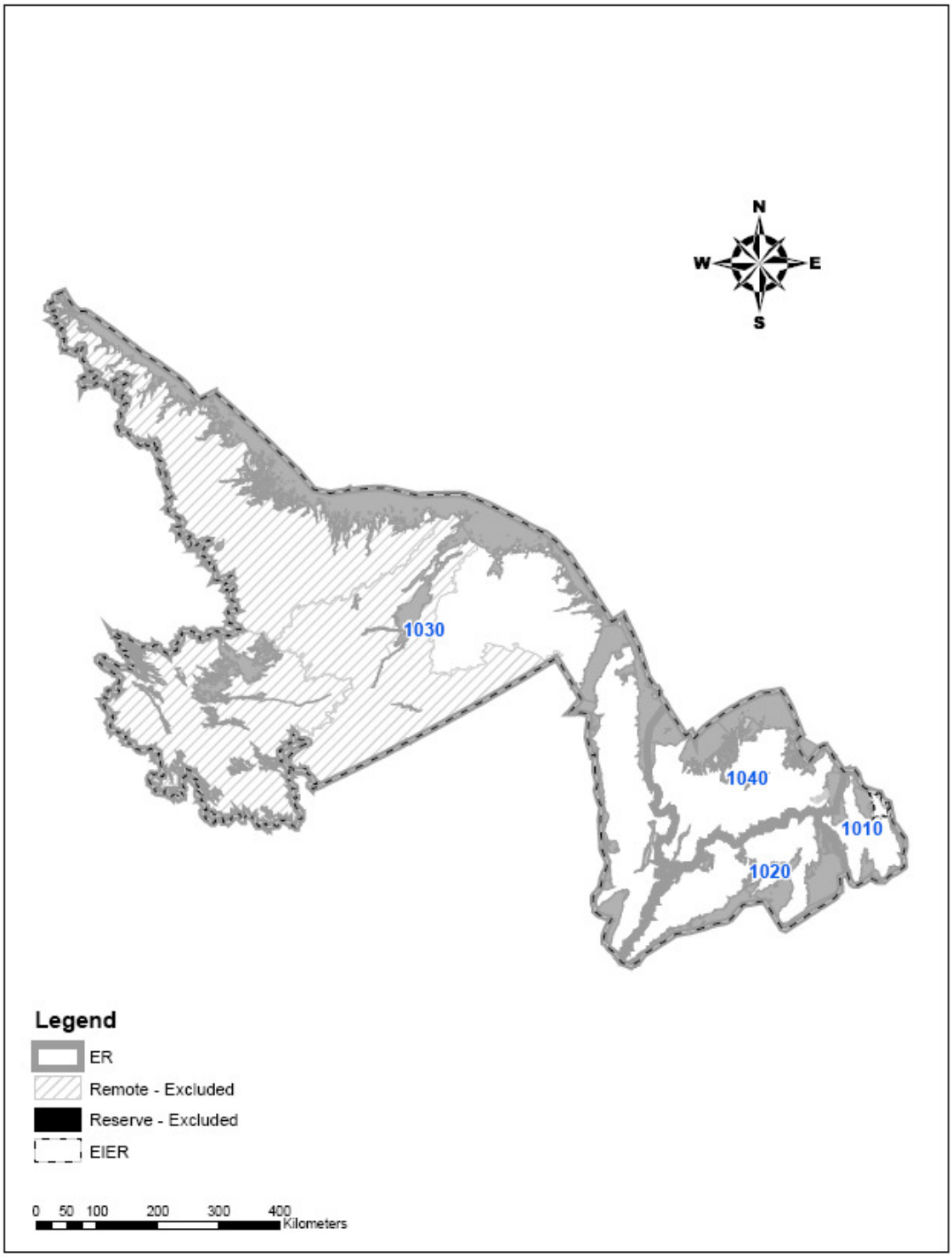


Appendix E Provincial Maps

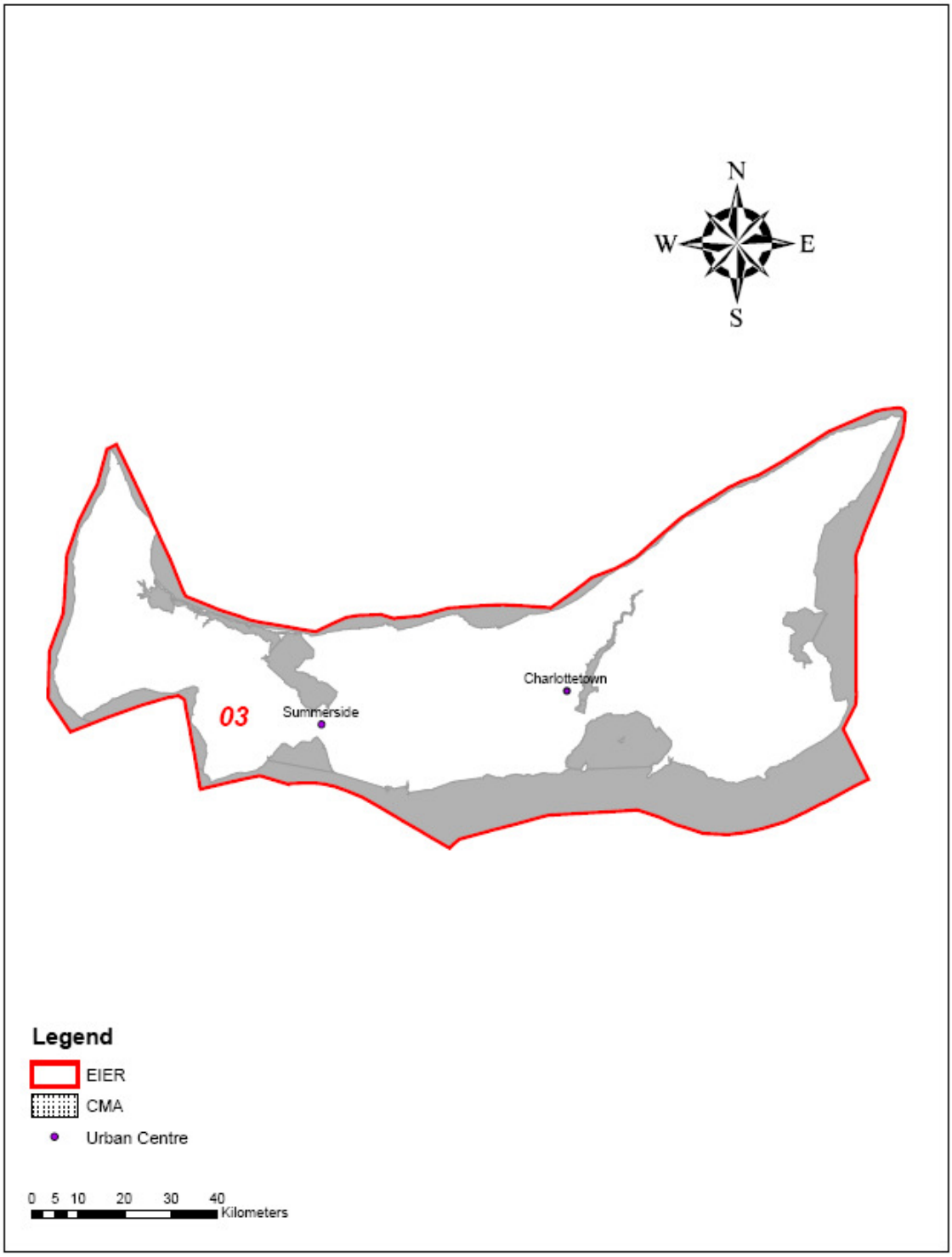
Map 1 Newfoundland and Labrador
EIERs and CMAs



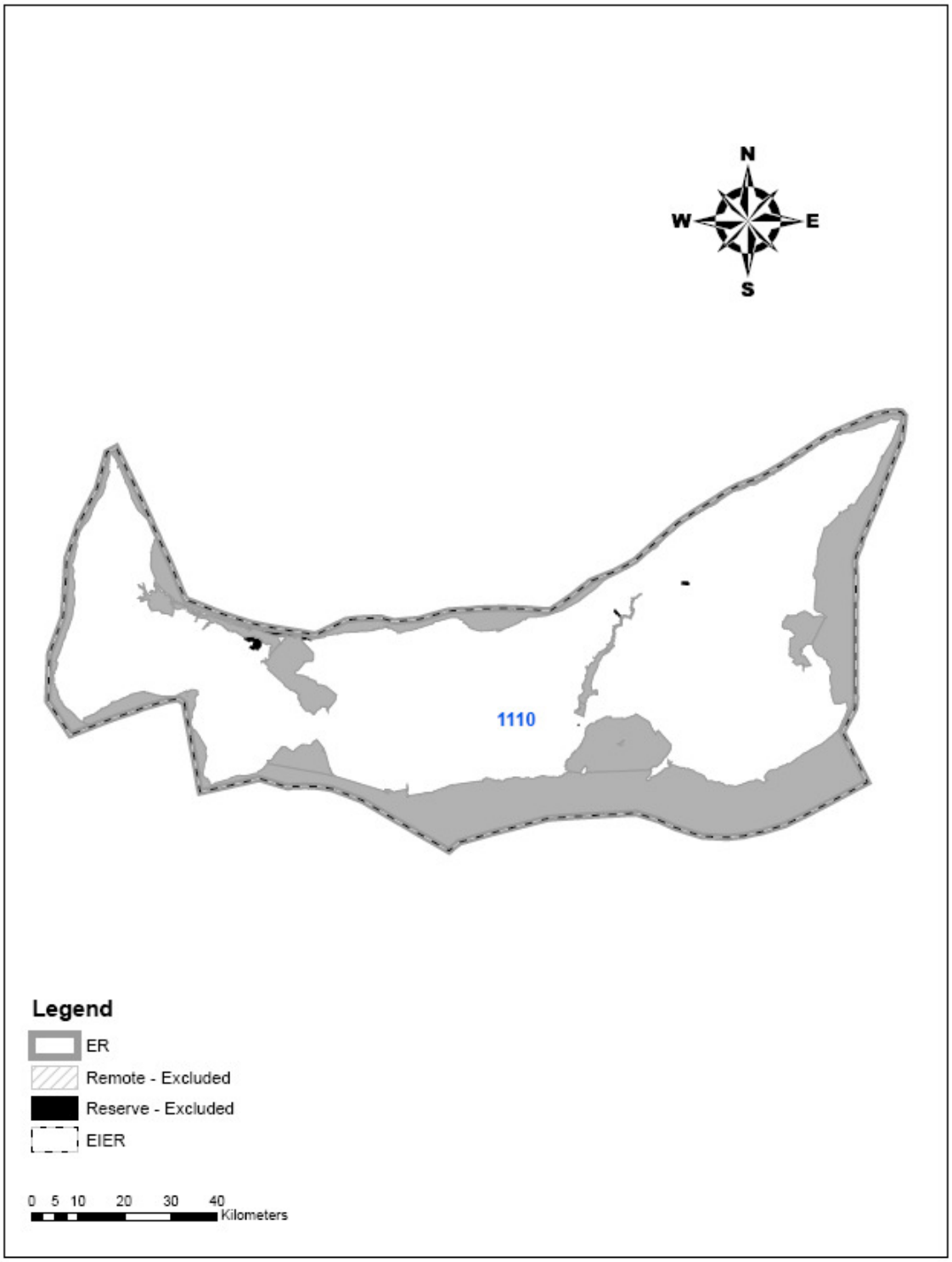
Map 2 Newfoundland and Labrador
Economic Regions



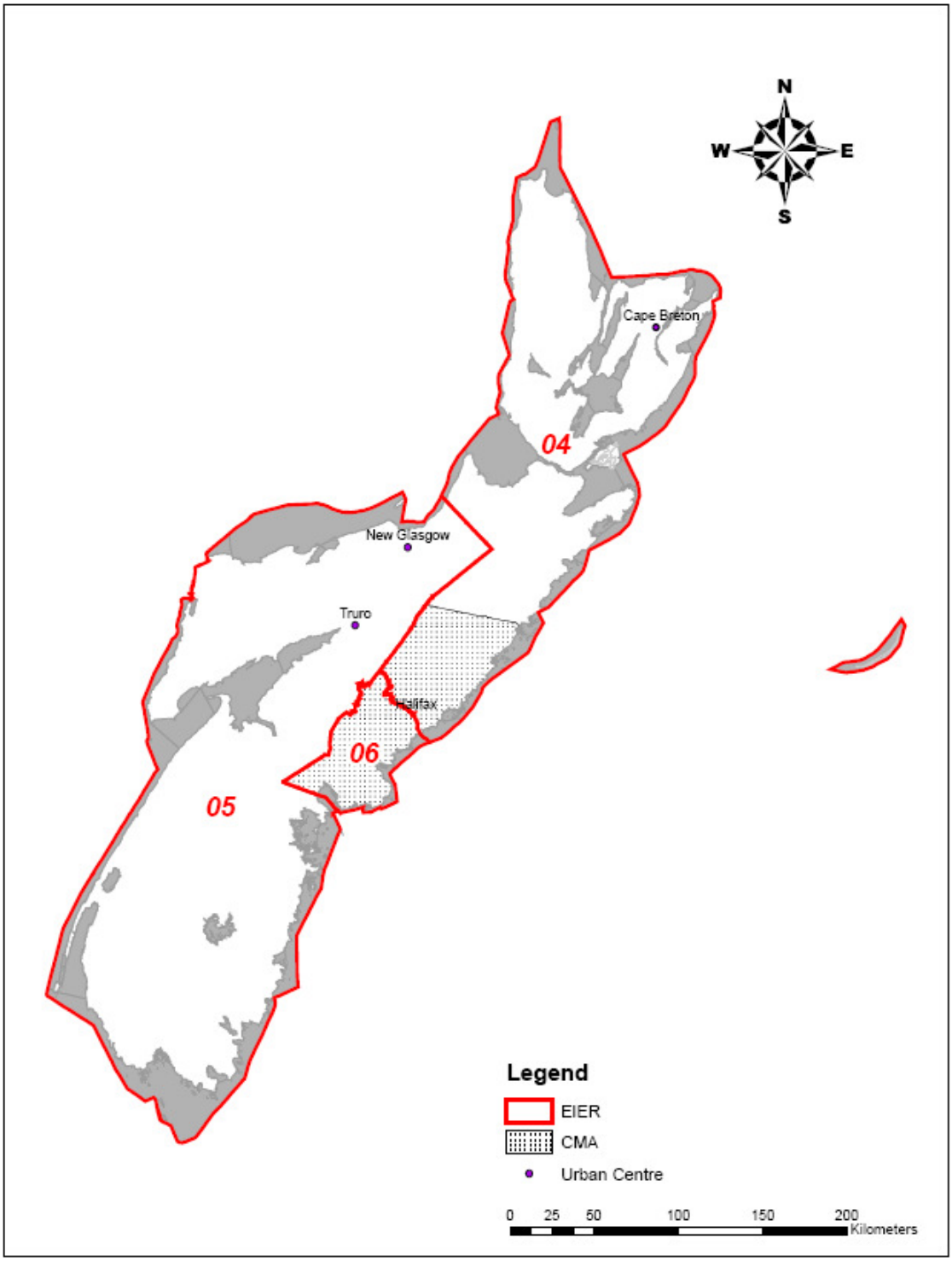
Map 3 Prince Edward Island
EIERs and CMAs



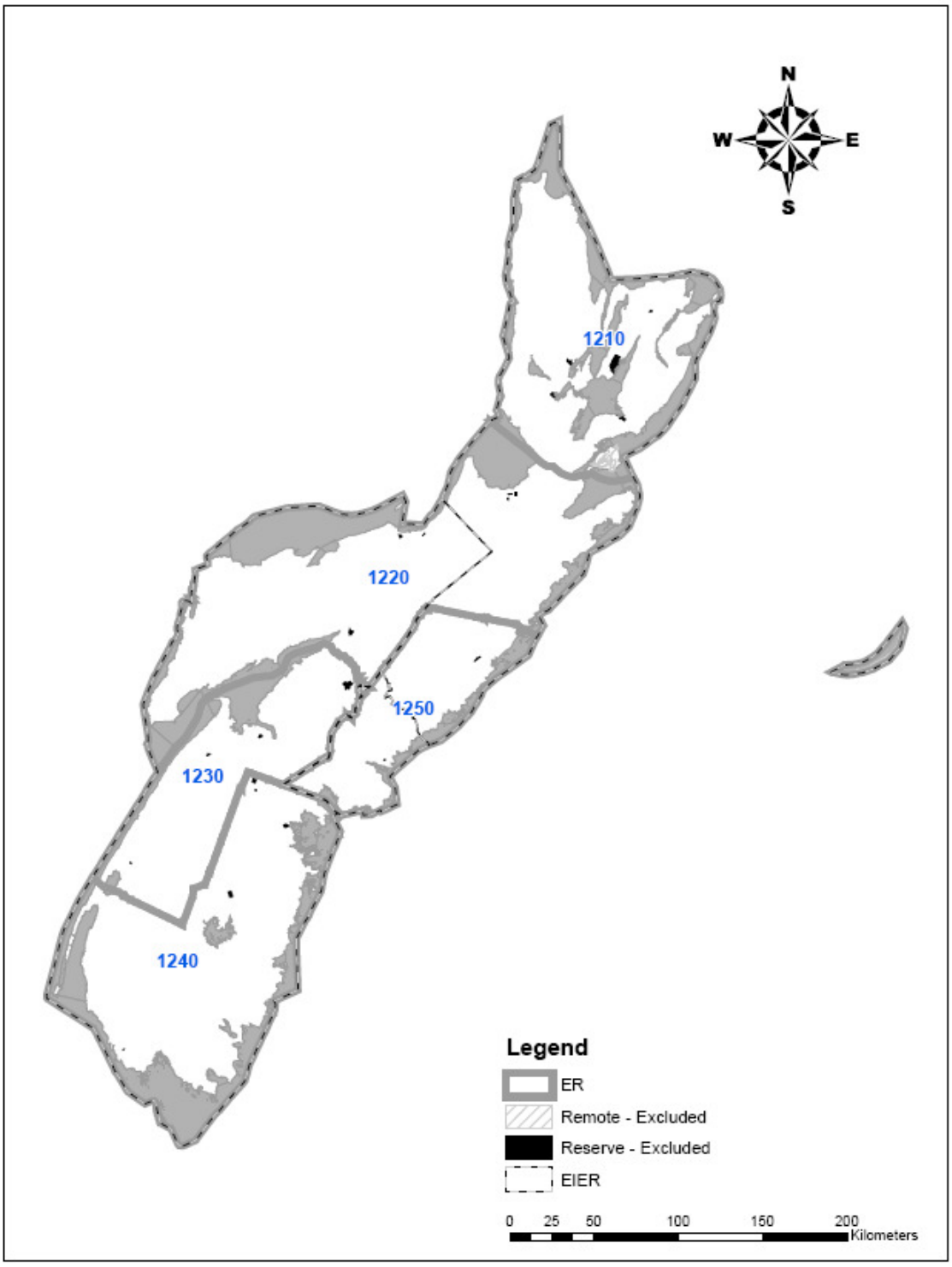
Map 4 Prince Edward Island
Economic Regions



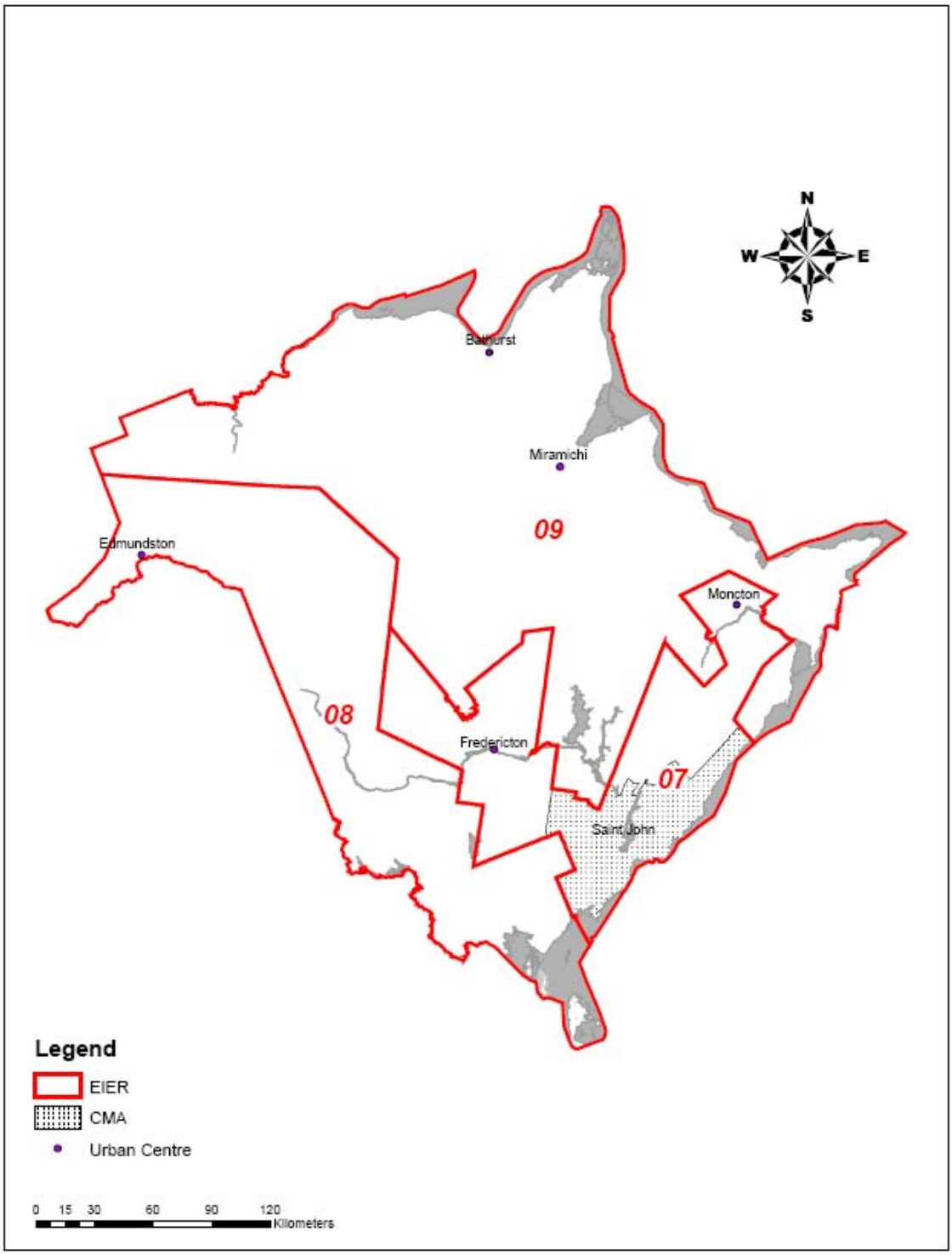
Map 5 Nova Scotia
EIERs and CMAs



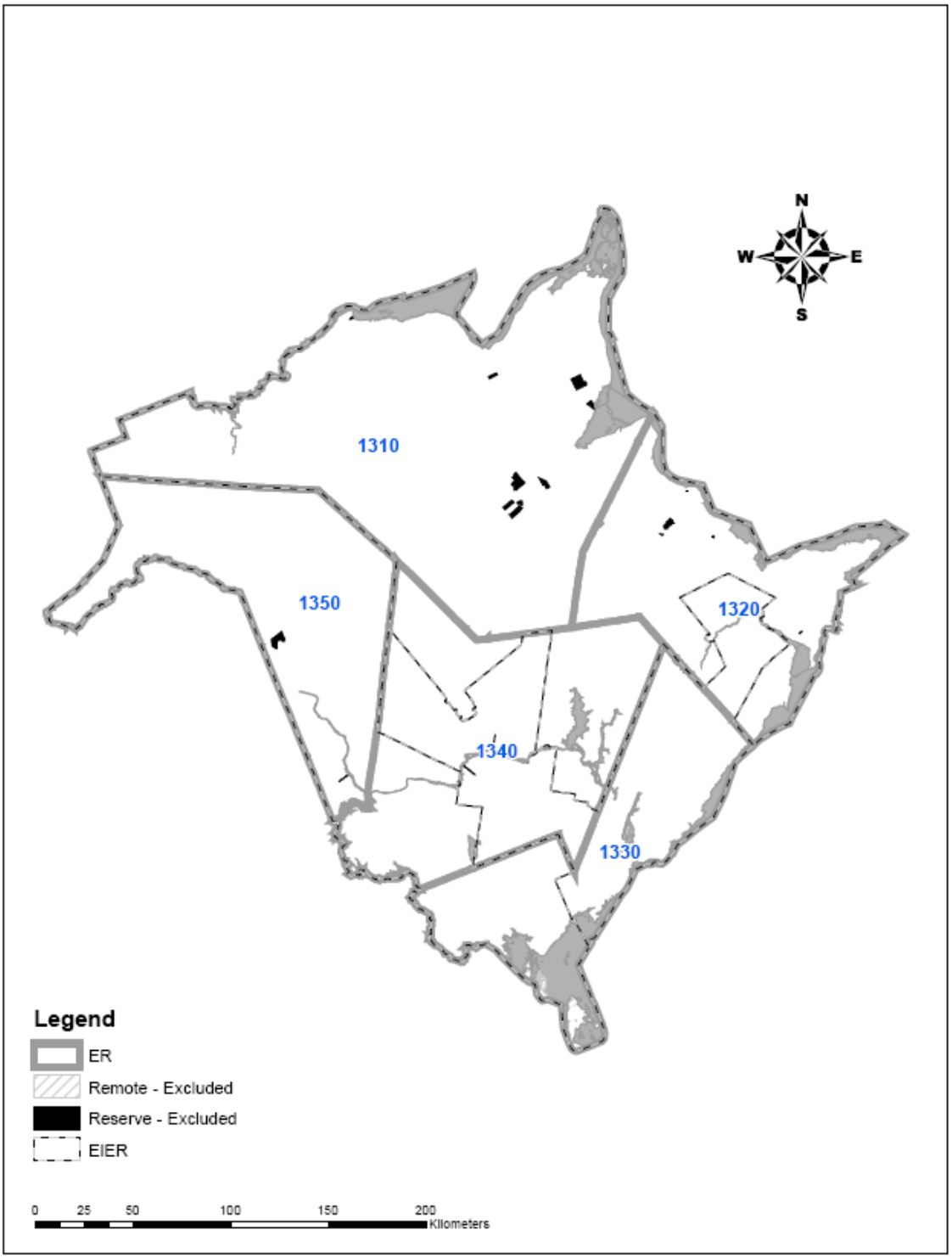
Map 6 Nova Scotia
Economic Regions



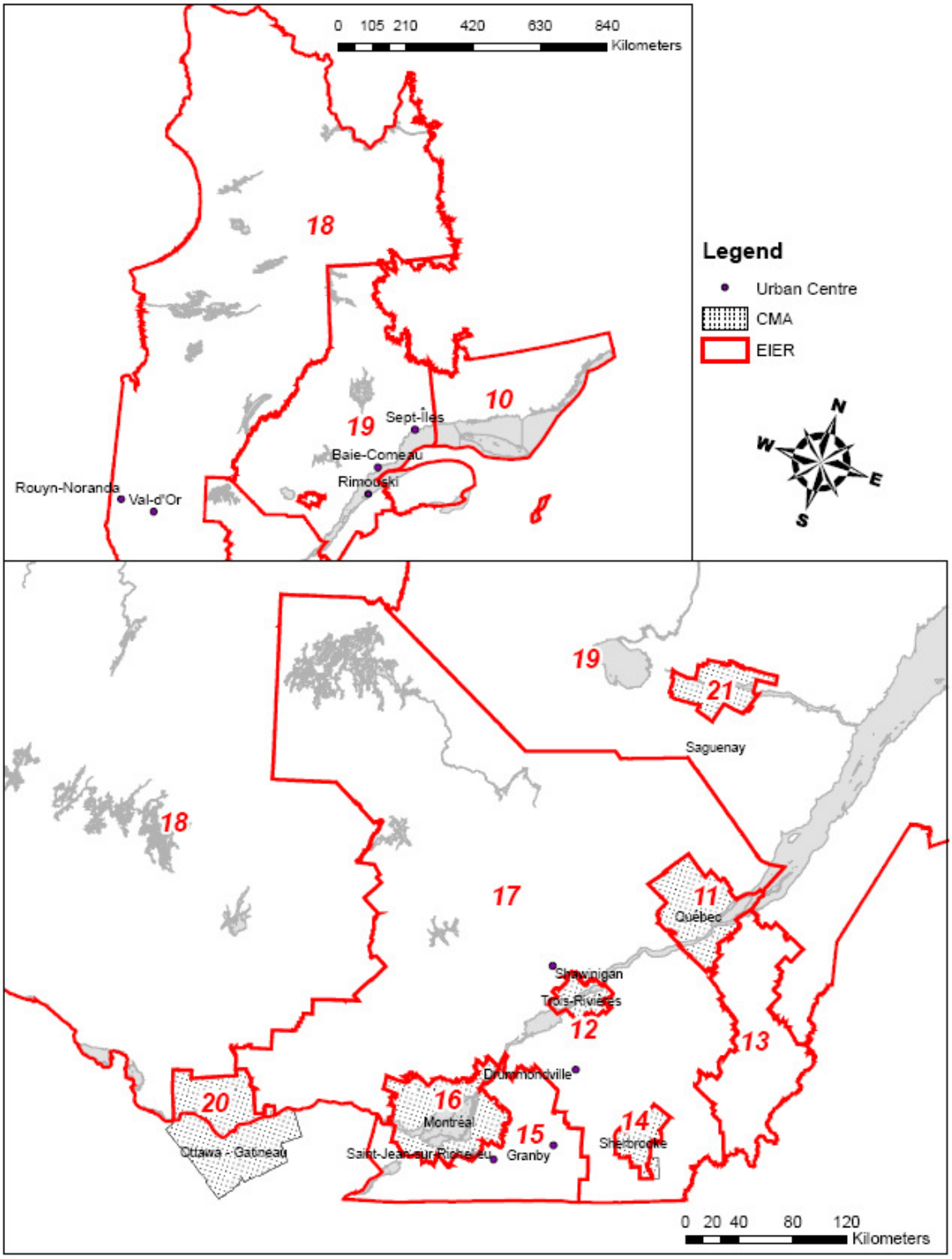
Map 7 New Brunswick
EIERs and CMAs



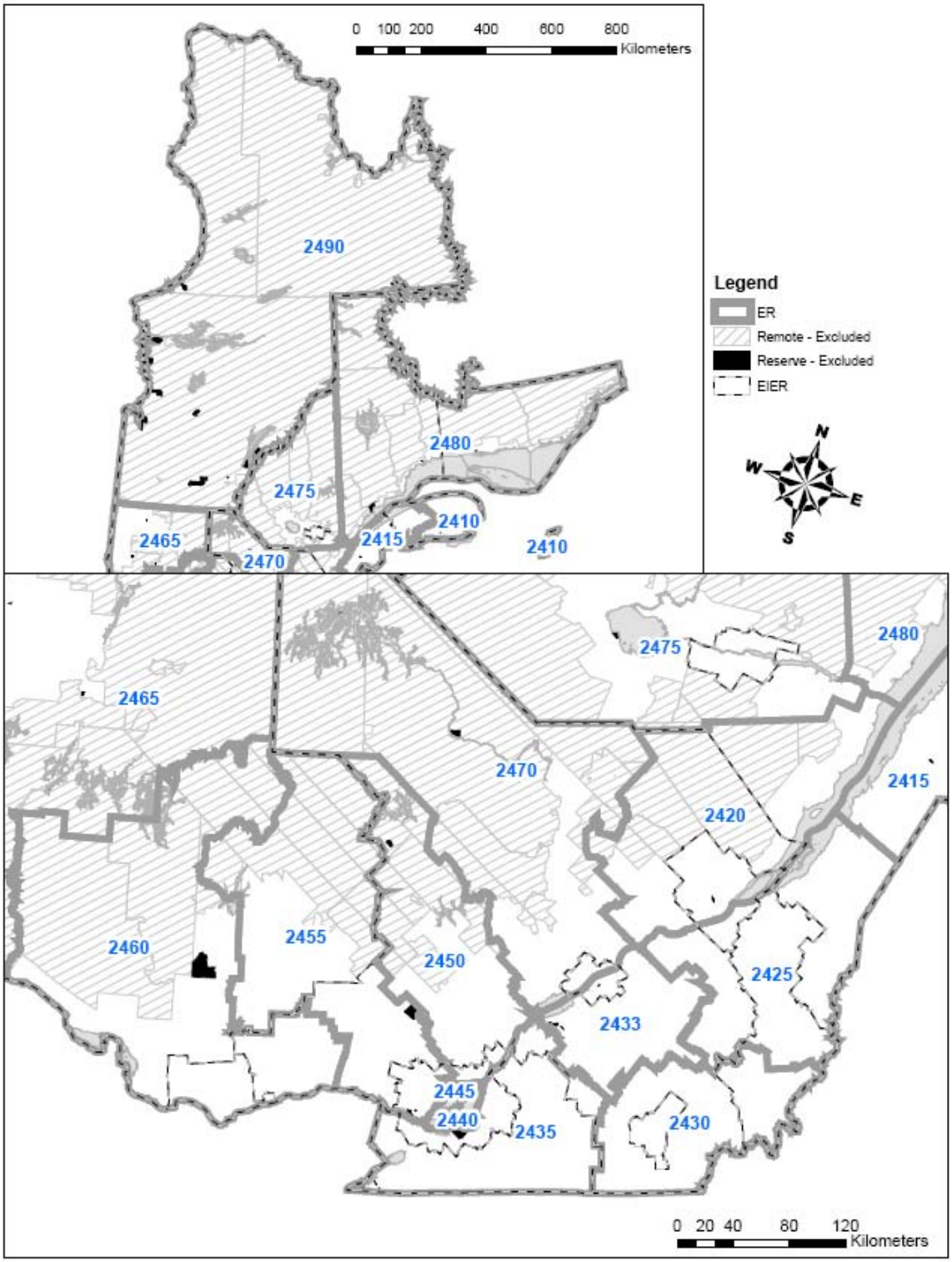
Map 8 New Brunswick
Economic Regions



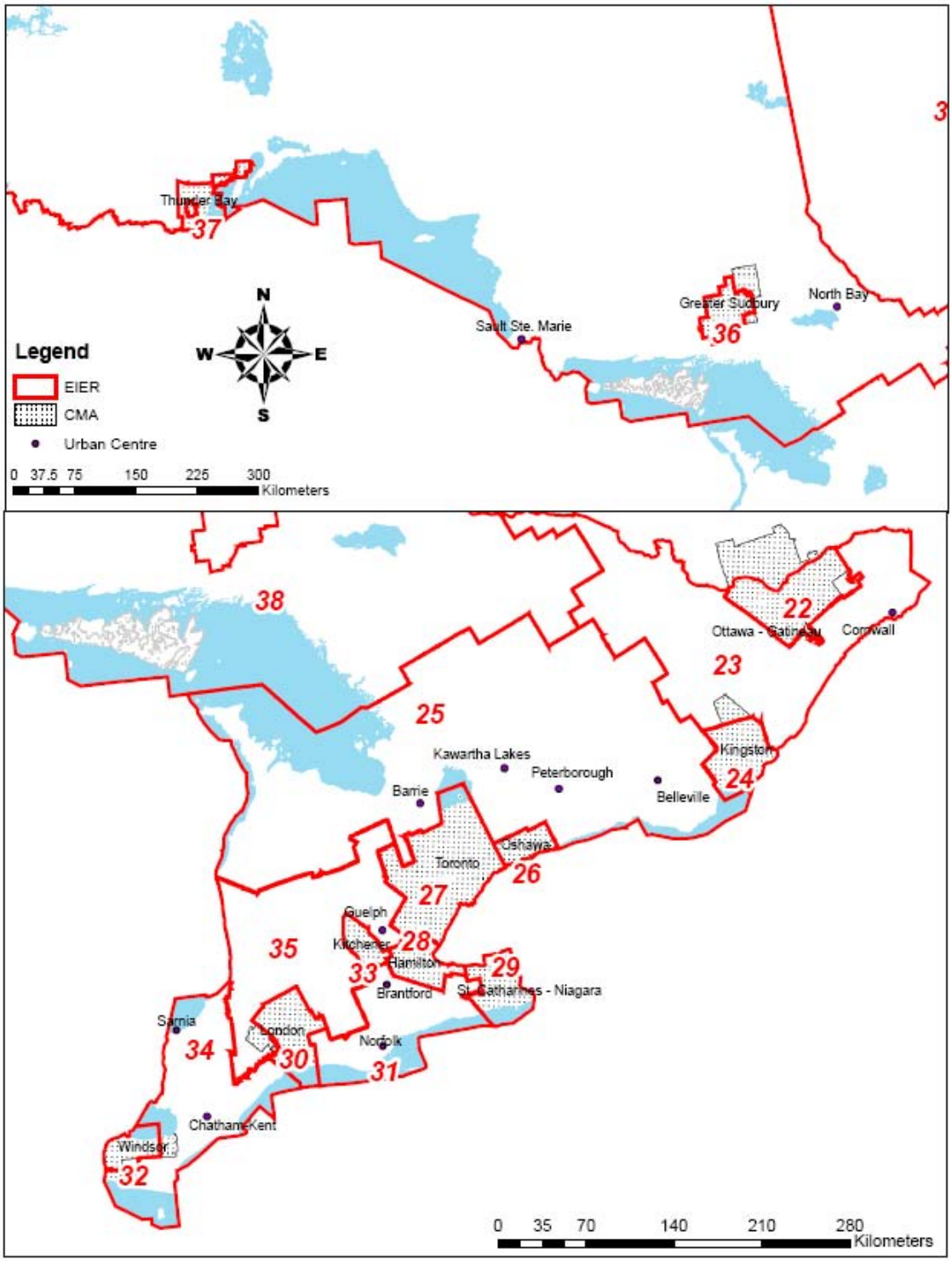
Map 9 Québec
EIERs and CMAs



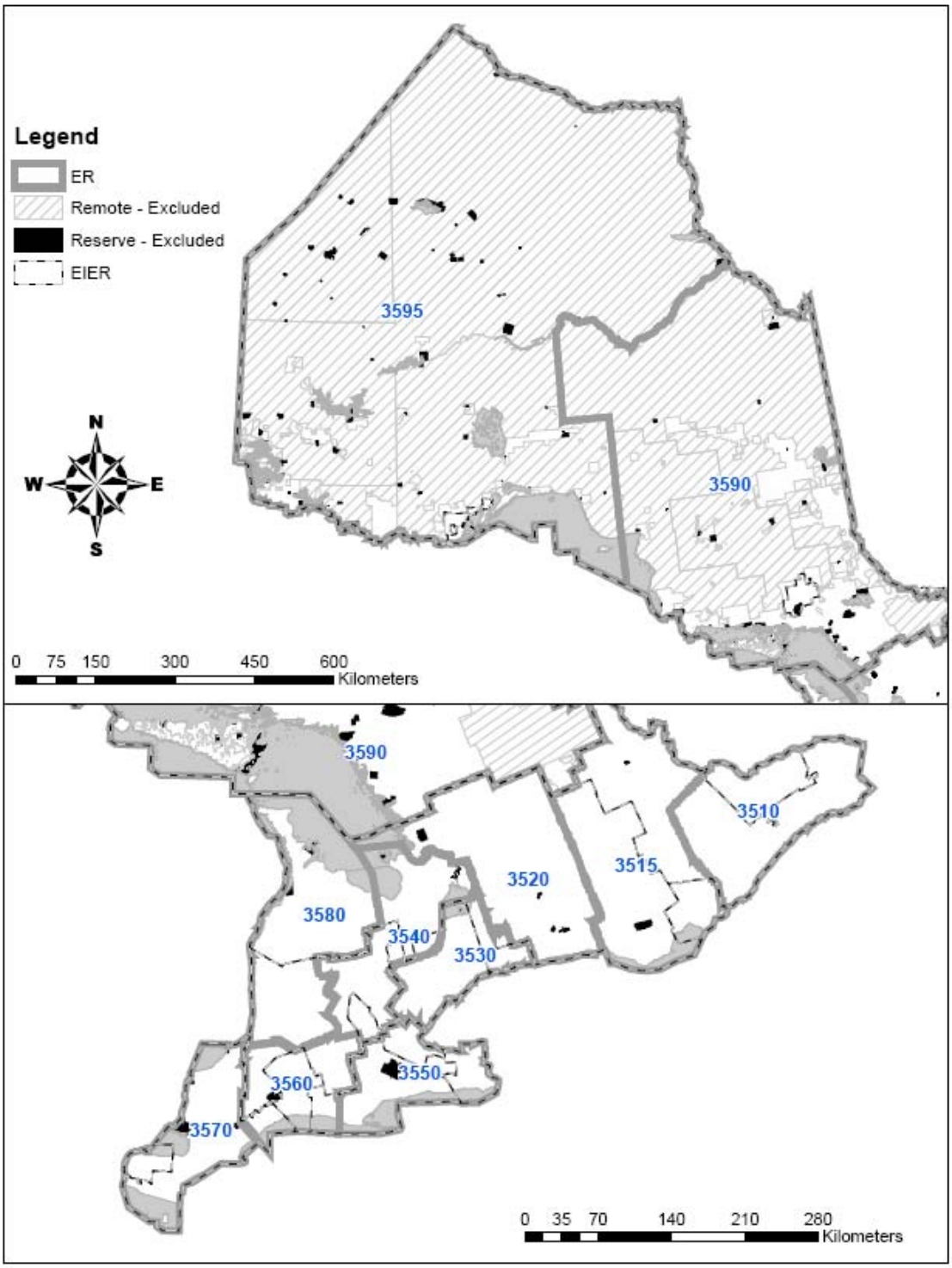
Map 10 Québec
Economic Regions



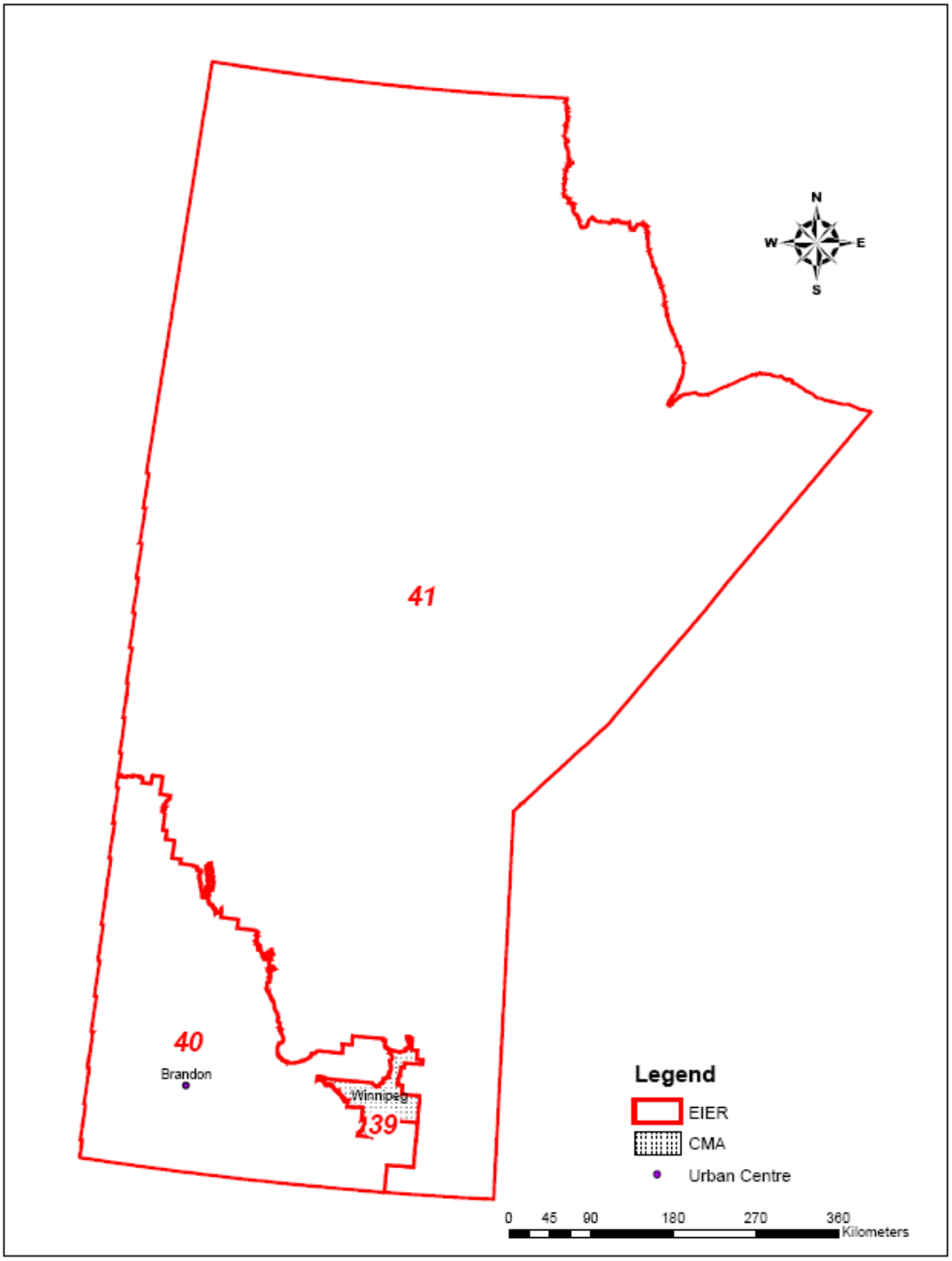
Map 11 Ontario
EIERs and CMAs



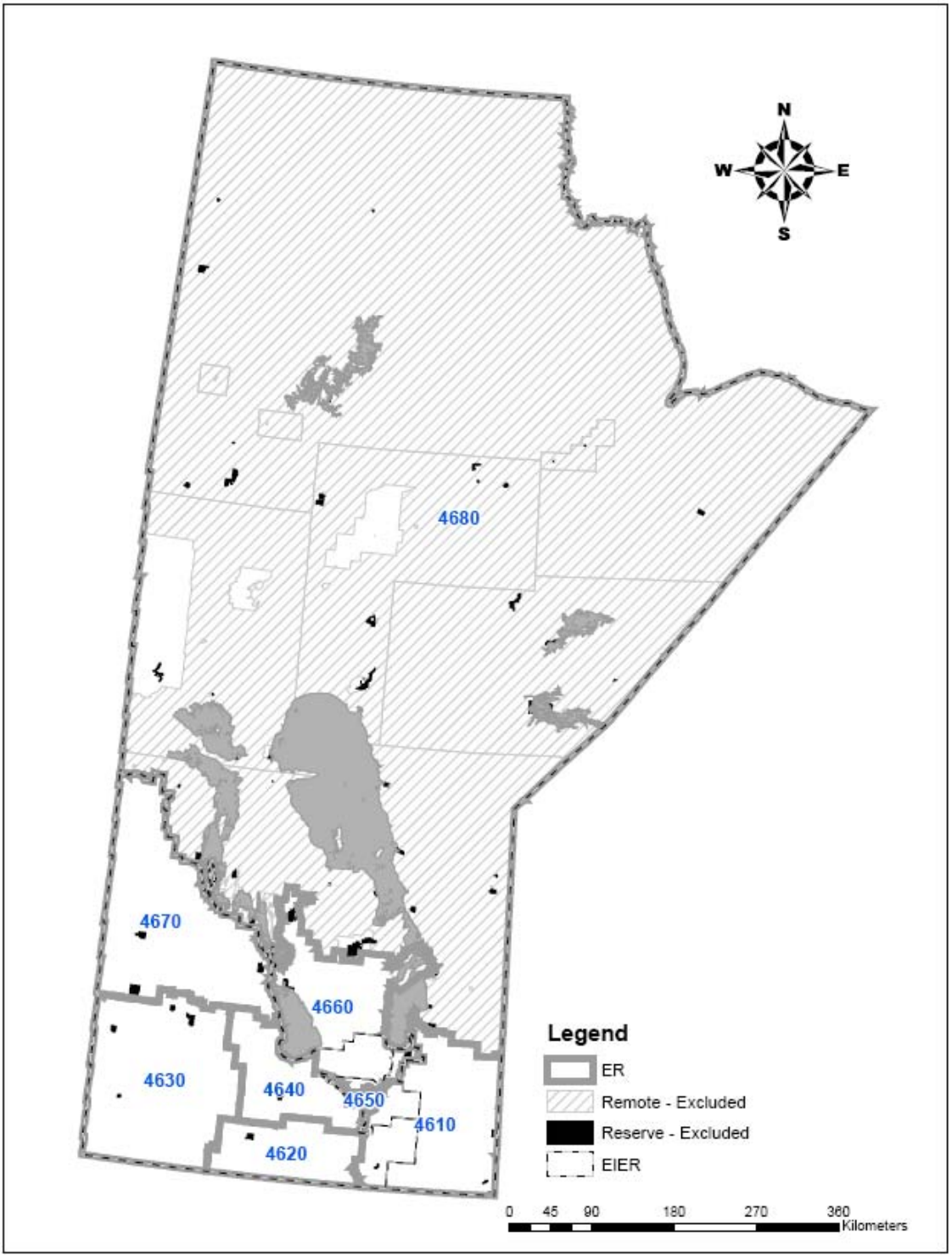
Map 12 Ontario
Economic Regions



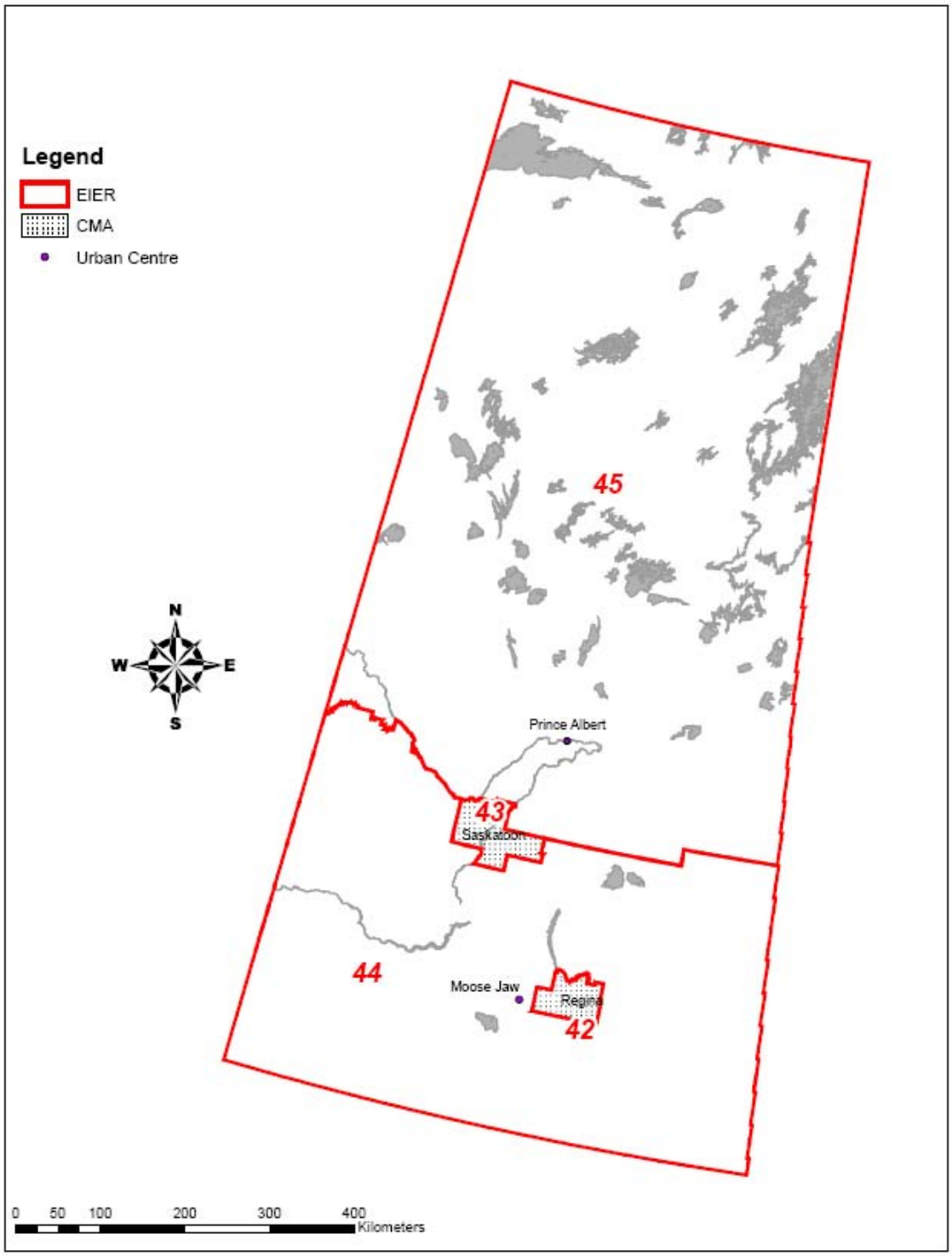
Map 13 Manitoba
EIERs and CMAs



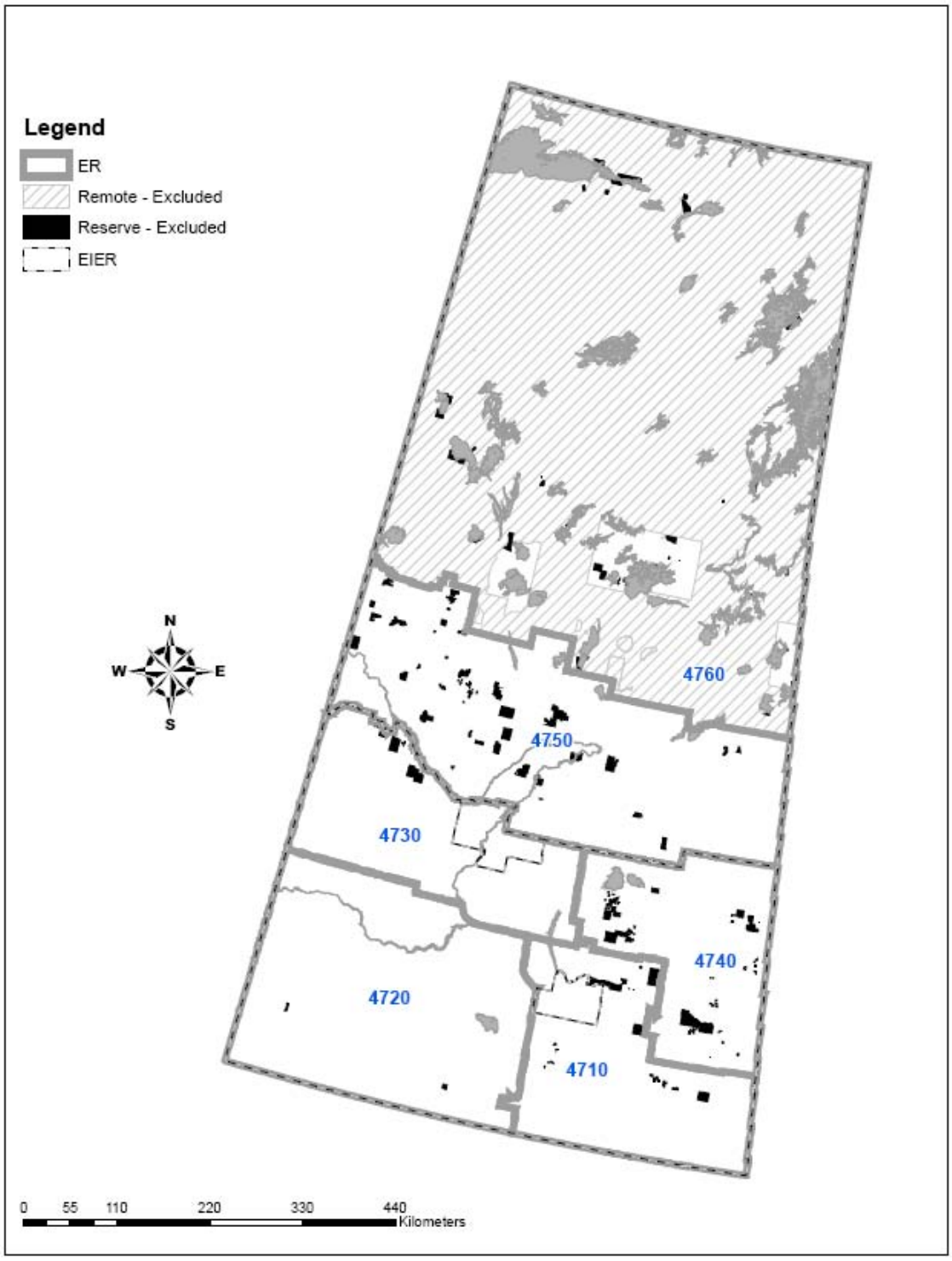
Map 14 Manitoba
Economic Regions



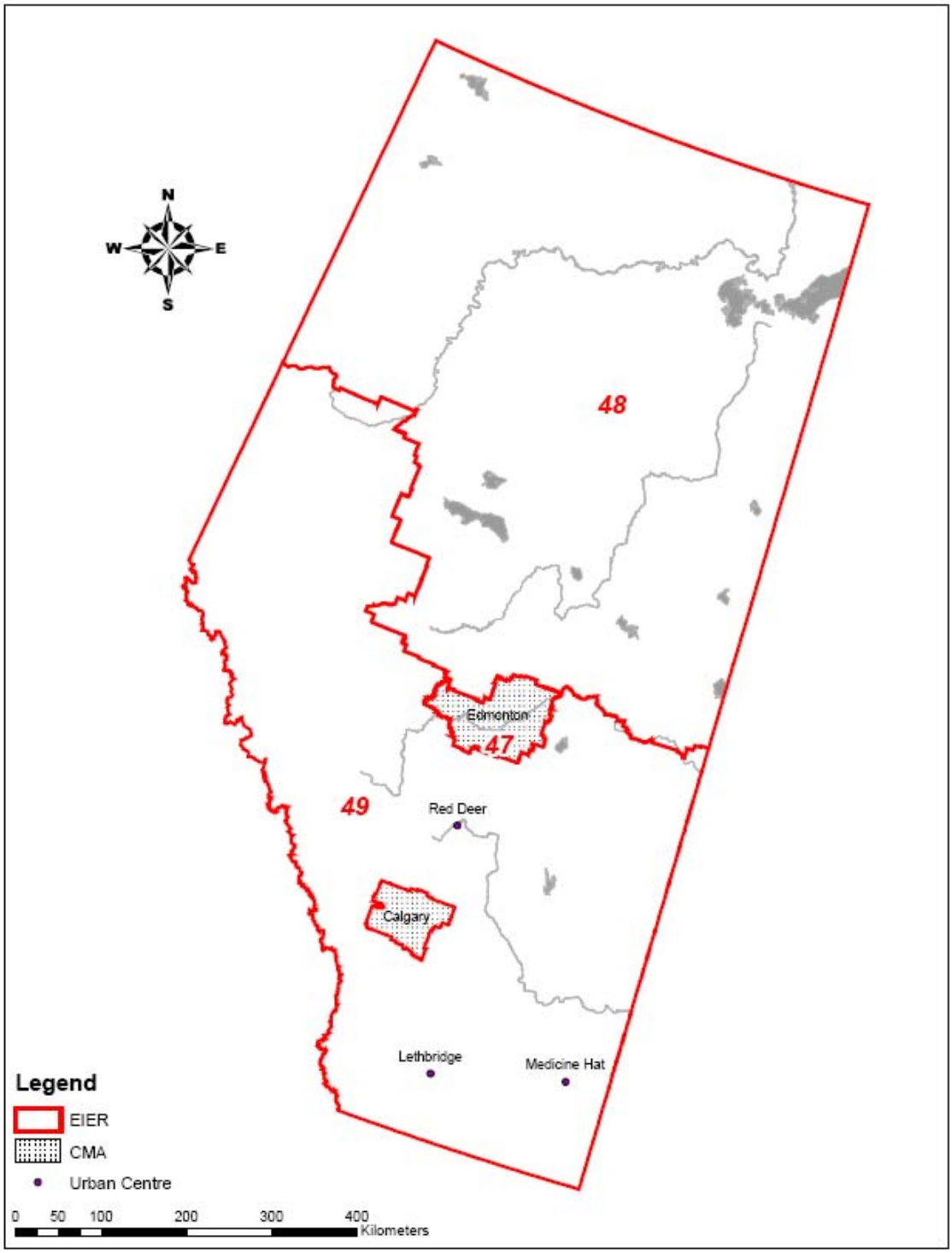
Map 15 Saskatchewan
EIERs and CMAs



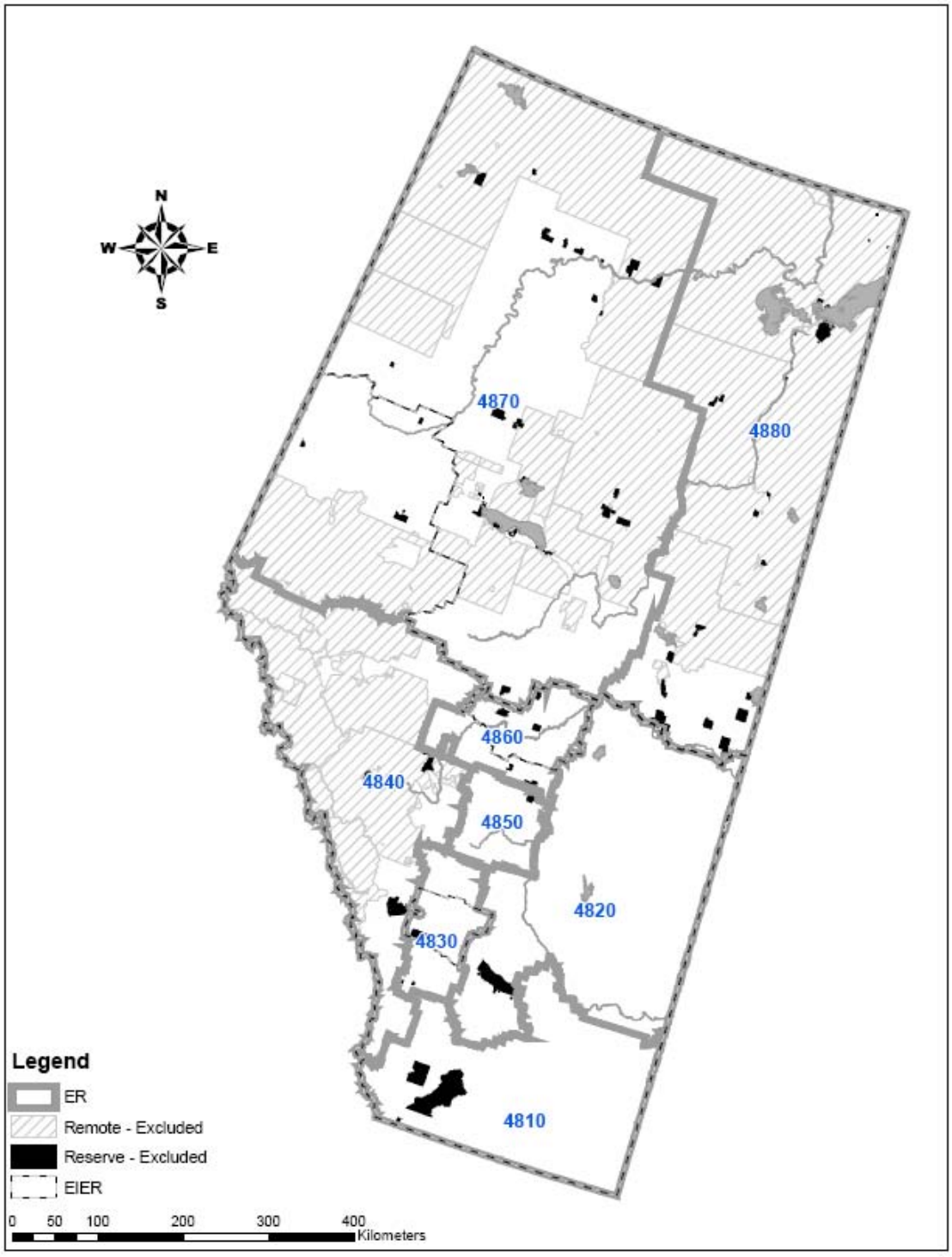
Map 16 Saskatchewan
Economic Regions



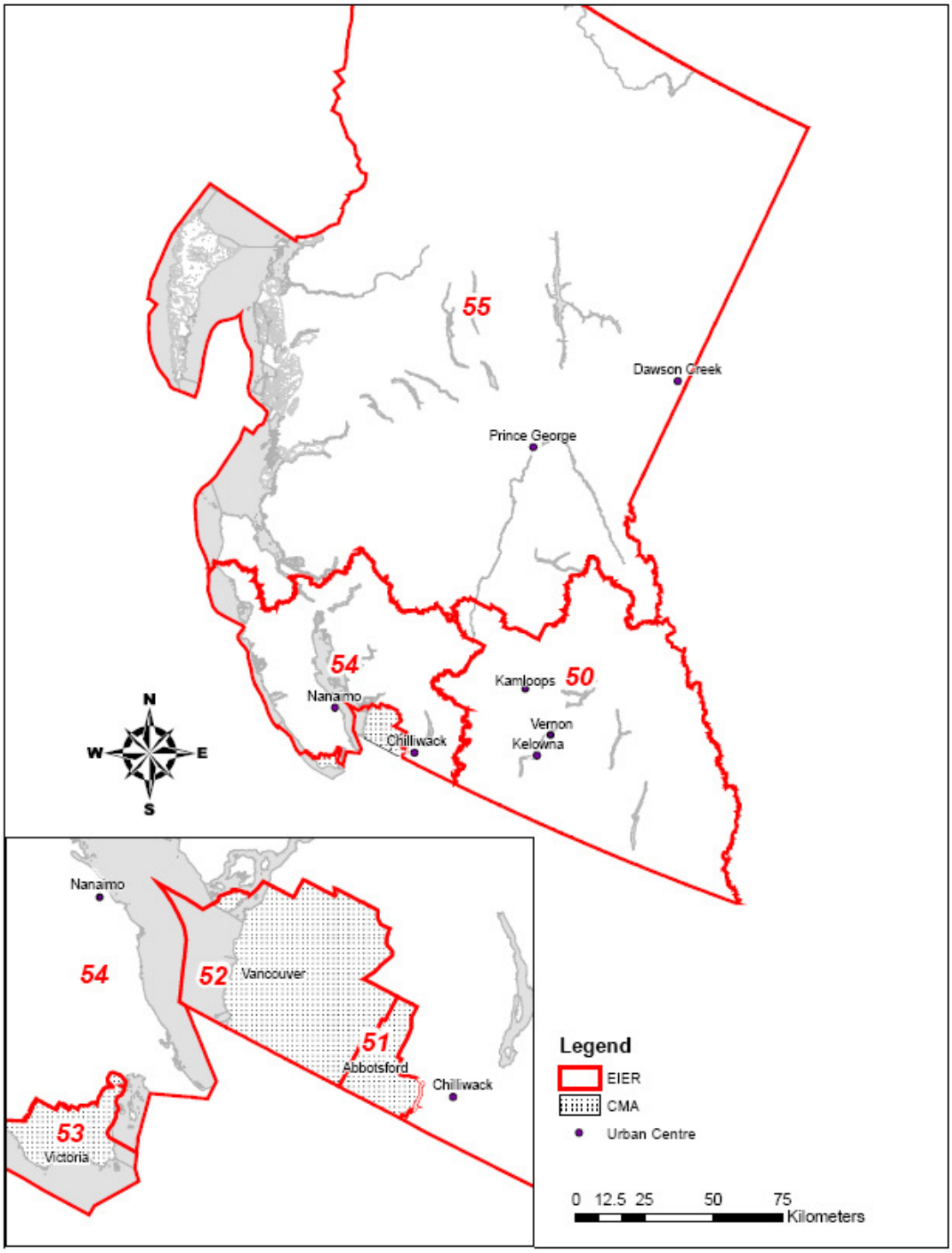
Map 17 Alberta
EIERs and CMAs



Map 18 Alberta
Economic Regions



Map 19 British Columbia
EIERs and CMAs



Map 20 British Columbia
Economic Regions