Catalogue no. 75F0002MIE — No. 002
ISSN: 1707-2840
ISBN: 0-662-40003-8

Research Paper

Income Research Paper Series

Recent changes in geography content in the Survey of Labour and Income Dynamics (SLID)

by Chris Li, Gaétan Garneau and Heather Lathe

Income Statistics Division
Jean Talon Building, Ottawa, K1A 0T6

Telephone: 613 951-7355

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-7355; (888) 297-7355; income@statcan.ca).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Depository Services Program inquiries	**1 800 700-1033**
Fax line for Depository Services Program	**1 800 889-9734**
E-mail inquiries	**infostats@statcan.ca**
Website	**www.statcan.ca**

Information to access the product

This product, catalogue no. 75F0002MIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that 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.

Statistics Canada
Income Statistics Division

Income research paper series

Recent changes in geography content in the Survey of Labour and Income Dynamics (SLID)

Published by authority of the Minister responsible for Statistics Canada

March 2005

Catalogue no. 75F0002MIE2005002

ISSN: 1707-2840

ISBN: 0-662-40003-8

Frequency: Irregular

Ottawa

La version française de cette publication est disponible sur demande (n° 75F0002MIF 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.

Abstract

This paper describes the changes made to the structure of geography information on SLID from reference year 1999 onwards. It goes into reasons for changing to the 2001 Census-based geography, shows how the overlap between the 1991 and 2001 Census-based concepts are handled, provides detail on how the geographic concepts are implemented, discusses a new imputation procedure and finishes with an illustration of the impact of these changes on selected tables.

Table of Contents

Introduction

Like many other household surveys, the Survey of Labour and Income Dynamics (SLID) asks its respondents to provide their current address every time they are interviewed. Address information is used not only for survey collection purposes, but also for assigning geographic information during data processing. Detailed geographic information permits data analysis for different types of regions within a province or across provinces.

The norms for coding geographic information in Statistics Canada are set by the Geography Division, primarily for the Census of Population, but also for the benefit of surveys. In preparation for the Census every five years, the Geography Division updates the geography-related lists of Canadian communities, to reflect any changes in their names or administrative boundaries that occurred since the previous Census. The new Census population counts are then used to categorize different areas according to their population density and other geographic characteristics.

Therefore, the SLID data always follow the geographic design of a particular Census. When the survey began collection in January 1993, the 1991 Census-based design was chosen, as it was the most recent one available. There was no update of SLID to the 1996 Census, because the need to change was not yet great enough. But starting with the publication of data for 2002, SLID changed its data processing to use the 2001 Census-based design. Both the cumulative geography changes since 1991 and the introduction of new concepts and methods by Geography Division in the 2001 Census design made it necessary to change the survey processing.

In making the decision to switch to the 2001 Census-based design starting with 2002, the SLID staff decided to ensure that not only the data for 2002 could be analysed using the new geographic design, but also the preceding three years, to provide a consistent four-year time series for the newer geographic detail. This period was a compromise between either revising none of the past years or revising all of them.

While the 1991 Census-based geography variables were maintained on the database, the new 2001 Census-based geography variables were added, with a one-to-one correspondence with the existing variables. It was also decided to maintain identical code sets for both sets of variables. To do this, the SLID staff modified and expanded the code sets to include codes for any new or re-named areas found in the new design, as compared to the 1991 design.

As for the published tables which include geographic detail below the provincial level (usually Census Metropolitan Areas and Census Agglomerations), they were updated so that the years 1999 to 2002 reflect the 2001 design. The estimates for years prior to 1999 were not revised and still reflect the 1991 Census-based design.

Finally, the SLID staff also took advantage of the change from the 1991 to the 2001 Census-based design to introduce improved methods for filling in missing values of geographic information, caused by missing address information. This process is called imputation. The improved imputation methods were also applied when the data for 1999, 2000 and 2001 were revised.

The postal code: the starting point for geography information

The SLID database contains several types of geographic detail. Although they are not all listed here (they are listed in Table 3: Geography variables on SLID database), some of the most important ones are: the Federal Electoral District (FED), dissemination area (DA), census metropolitan area/census agglomeration (CMA/CA), census division (CD), census subdivision (CSD), economic region (ER), latitude and longitude of residence, and employment insurance region (EIR).[1]

The building block for all these types of areas is the postal code. The area represented by a postal code is usually small enough, at least in more urbanized areas, to determine unambiguously all other geographical areas of interest. The postal code is obtained as part of the respondent’s address, each time they are interviewed. Fortunately, most respondents know their postal code. Nevertheless, for some households in the sample, it is “missing” when it comes back from collection, due to non-response or a data capture error. For these records, the postal code is derived via the telephone number or it is imputed during survey processing in order to have a complete file. Only then are all the rest of the geographical variables derived.[2]

Reasons for changing to 2001 Census-based design

Certain geographic classifications need to be updated with every Census. First of all, the categorization of communities into Census Metropolitan Areas (CMAs), Census Agglomerations (CAs) or non-CMA/CA areas depends on their population as determined by the Census of Population, which is conducted every five years. Secondly, Geography Division changes the boundaries from one census to another to include new neighbourhoods or exclude others, the corresponding population changes, and this alone may change the profile of people living there. If the data are not updated to reflect the boundary changes, then they may not be adequately representative of the new area.

Up until reference year 2001, the geographic information on SLID was based on the 1991 Census-based design. Starting with the release of data for reference year 2002, the geographical information of SLID is based on the 2001 Census-based design.

[1] For more information on these concepts, please refer to: Census Operations Division, 2002, “2001 Census Dictionary, Reference”, Statistics Canada, Catalogue no. 92-378-XPE.

[2] There is one drawback associated with the postal code as the sole collection input for geographic coding: it does not always represent the residence of the household. This is particularly the case in rural areas, where rural route service and post office pick-up is commonly used to deliver mail. But, in the majority of cases, the postal code corresponds to the location of the respondent’s residence.

There were a few reasons why SLID did not change to the 1996 Census-based design when it became available. Several tests performed on the SLID data showed that the changes from the 1991 to the 1996 Census-based design were not important enough to warrant a change in the SLID geography coding at that time. However, when the 2001 Census-based design became available, it became clear that a change was necessary. Not only were there cumulative changes in population counts over a ten-year period, but also there were many boundary changes throughout the second half of the decade (1996 to 2001)—somewhat more than in the first part of the decade.

To illustrate this second effect, Table 1 below shows the impact of administrative boundary changes on the population of fifteen major CMAs, for the years 1991 to 1996, 1996 to 2001, and the combined ten-year period from 1991 to 2001.

Between 1991 and 1996, there were relatively small changes in population due to boundary adjustments. Out of the 15 CMAs published in SLID's standard tables, only three of them had boundary changes that caused the population estimate to change by one percent or more. Between 1996 and 2001, only four had such significant boundary changes. However, six of the fifteen CMAs were affected between 1991 and 2001, suggesting a change was necessary.

Table 1: Percentage change in population due to boundary changes of CMAs

CMA Name	Population change due to boundary changes between Censuses (%)		
	1991 to 1996	1996 to 2001	1991 to 2001
Montreal	2.7%	0.0%	2.8%
Ottawa-Gatineau	2.4%	-1.2%	1.3%
Winnipeg	1.3%	0.0%	1.3%
Halifax	0.0%	3.3%	3.5%
Québec City	0.0%	0.0%	0.0%
Toronto	0.2%	0.0%	0.2%
Hamilton	0.0%	0.0%	0.0%
St. Catherines-Niagara	0.0%	0.0%	0.0%
Kitchener	0.0%	0.0%	0.0%
London	0.0%	4.7%	4.8%
Windsor	0.0%	3.1%	2.9%
Calgary	0.0%	0.0%	0.0%
Edmonton	0.1%	0.0%	0.2%
Vancouver	0.0%	0.0%	0.0%
Victoria	0.0%	0.0%	0.0%

Notes:

The percentages in the table may not sum up due to rounding.

The 0.0% values in the table implies that there was either no boundary change between the Censuses or the population did not change significantly even though there was a boundary change.

Source: GeoRef 91 (Catalogue no. 92-345D); GeoSuite 2001 (Catalogue no. 92F0085XCB); and Geography Division.

Overlap of the Census-based geography designs in SLID data

In making the switch to using the 2001 Census-based design, the SLID staff decided to historically revise the 1999, 2000 and 2001 sub-provincial data to follow the geographic boundaries of the 2001 Census-based design, while also keeping the earlier values in a separate set of variables. It gives an overlap of three years that can be studied using either the 1991 or 2001 Census geographic boundaries. This choice makes data analysis possible using a uniform geography design for the full six years of each panel of respondents. Table 2 shows how the database is structured.

Table 2: Structure of the geography variables

Reference years:	1992 to 1998	1999 to 2001	2002 on
Variables without ‘X’ e.g. CMACA25	Refers to 1991 boundaries	Refers to 2001 boundaries	
Variables with ‘X’ e.g. XCMACA25	Refers to 1991 boundaries		Coded to ‘Not Applicable’

All the 1991 Census-based design geography values, covering years 1992 to 2001, are stored in variables that have the same names except they begin with an ‘X’. For example, data for CMACA25 was copied to XCMACA25. The regular geography variables contain the 2001 Census-based design values, for the period 1999 onwards.

Most of the geography variables refer to the household’s location at the end of the reference year. But there are four variables on geographic mobility that are directly affected by the change to the 2001 Census-based design: SAMCMA26 – for those living in a CMA/CA at then end of a reference year, this is a flag indicating whether the person is living in the same CMA/CA as at the end of the previous year; SAMCSD26 – flag to indicate whether the person is living in the same CSD as at the end of the previous year; DISTM26 – distance moved expressed in kilometres; and MOVST29 – move status for the person for the reference year.

These four variables are derived from the other geography variables by comparing the previous and current reference years’ values. Since two years of data are required to derive a value for a given year, they now use the 2001 Census-based design starting with 2000 (which is the comparison of 1999 and 2000). For reference years up to and including 1999, they use the 1991 Census-based design.

New and discontinued standard geographic concepts

The 2001 Census-based design includes three new standard geographic concepts that were not present in either the 1991 or 1996 Census-based designs (or earlier ones).

1. The **dissemination area (DA)**[3] is a small, relatively stable geographic unit composed of one or more blocks. Starting with 2001, the dissemination area is the smallest standard geographic area for which any census data can be disseminated (i.e., tabulated for analysis), hence the name chosen for it. Previously, the smallest area for disseminating data was the **enumeration area (EA)**[4]. The EA is a geographic area canvassed by one census representative, and it is still used by the Census for collection purposes.

2. The **census metropolitan area and census agglomeration influenced zone (MIZ)**[5] is a new concept introduced starting with the 2001 Census-based design. It classifies Census sub-divisions (CSDs) that are not part of either a CMA or CA according to the degree of influence that CMAs and/or CAs collectively have on them. The degree of influence on the CSD is estimated by the extent of the commuting flow between it and other areas—specifically, the proportion of the total employed labour force living in the CSD that works in any CMA/CA urban core. The CSDs with a commuting flow of 30% or more are classified as the "strong" MIZ. The CSDs with a commuting flow of at least 5% and less than 30% are called the "moderate" MIZ. The CSDs with a commuting flow of more than 0%, but less than 5%, are included in the "weak" MIZ category. Those CSDs with fewer than 40 people in the resident labour force commuting to work in CMA/CA urban cores are classified as "no" MIZ.

3. The **Statistical Area Classification (SAC)**[6] groups census subdivisions (CSDs) according to whether they are a component of a CMA, CA, MIZ or the territories. This more detailed classification replaces the **CMA/CA type**. CMAs and CAs are no longer classified as either "consolidated" or "primary". Also, CAs (but not CMAs) are classified as "tracted" or "non-tracted". Finally, all other areas, which did not have any classification before, are now classified using the MIZ.

Table 3 contains the geography variables on the SLID database. The first two columns list the geography variables available on SLID with reference to the 1991 Census-based design. The third column indicates whether the same 1991 Census-based design variables are also available in the 2001 Census-based design.

[3] For more information on these concepts, please refer to: Census Operations Division, 2002, "2001 Census Dictionary, Reference", Statistics Canada, Catalogue no. 92-378-XPE.

[4] For more information on these concepts, please refer to: Census Operations Division, 1992, "1991 Census Dictionary, Reference", Statistics Canada, Catalogue no. 92-301E.

[5] For more information on these concepts, please refer to: Census Operations Division, 2002, "2001 Census Dictionary, Reference", Statistics Canada, Catalogue no. 92-378-XPE.

[6] For more information on these concepts, please refer to: Census Operations Division, 2002, "2001 Census Dictionary, Reference", Statistics Canada, Catalogue no. 92-378-XPE.

Table 3: Geography variables on SLID database

1991 Variable	Description	In 2001 Census-based design
PVRES25	Province of residence	Yes
PVREG25	Province of residence group	Yes
FEDRES25	Federal electoral district[7]	Yes
EARES25	Enumeration area	No, replaced by dissemination area: DARES25
CDRES25	Census division	Yes
CSDRES25	Census subdivision	Yes
REGRES25	Region	Yes
ERRES25	Economic region	Yes
CMACA25	CMA/CA	Yes
CMA1G25	CMA/CA group 1	Yes
CMA2G25	CMA/CA group 2	Yes
CMAPO25	CMA/CA population	Yes
CATYP25	CMA/CA type	No, replaced by standard area classification type: SACTYP25 (includes MIZ classifications)
LNGRES25	Longitude of residence	Yes
LATRES25	Latitude of residence	Yes
URBRUR25	Urban/rural household	Yes
URBCD25	Urban area code	Yes
URBPO25	Population of area of residence	Yes
URBSZ25	Urban size	Yes
URBSZG25	Size of area of residence	Yes
USZGA25	Adjusted size of area of residence	Yes
EIR25	Employment Insurance Region[8]	Yes
POSTCD25	Postal code	Yes
POSTCI25	Postal code imputation flag	Yes

Blending of two geography designs in a single set of geography variables

Those geography concepts which are common to both the 1991 and 2001 Censuses continue to exist under the same variable names in SLID as before the 2002 data release. Instead of creating new variables, the SLID staff augmented the categories and descriptions as necessary to incorporate new areas and names.

Any geography categories that only apply to certain years (because they only apply in one of the two Census geography designs) will still appear in the code set, but with a note saying for what years it is not applicable. There will be no households with those codes in those years.

[7] FEDs are represented by a member of the House of Commons.

[8] For the 1991 Census-based design, the July 1996 delineation of employment insurance regions is used. For the 2001 Census-based design, the July 2000 delineation of employment insurance regions is used.

Below are three examples that show how the variable CMACA25 has been modified to account for the two geography designs. Note that, in order to qualify as a Census Metropolitan Area (CMA) or Census Agglomeration (CA), a municipality must satisfy certain population criteria; CAs can be created or retired between Census designs, and although CMAs are never retired, there could be new ones created. Households in non-CMA/CA areas are given a value of “not applicable” for variable CMACA25.

Case A: Codes only applicable prior to 1999 (the 2001 Census-based design). The CMACA25 code 730 represents Weyburn, Saskatchewan. It was a CA in the 1991 Census-based design, but was retired in 1996 because the population fell below 10,000 according to the Census results (it retained its non-CMA/CA status in 2001). Households in Weyburn are no longer part of a CMA/CA, so they would be given the “not applicable” code for this variable starting in 1999. An example of a CMA/CA that disappeared because of a boundary change is Saint-Jérôme, Quebec, code 475. It was a CA on its own in the 1991 Census-based design, but by the time of the 2001 Census it had become amalgamated with Montreal, which is already its own CMA. Unlike Weyburn, the households in Saint-Jérôme are still part of a CMA/CA and would be given the code for Montréal starting in 1999.

Case B: Codes only applicable for years 1999 and on (the 2001 Census-based design). An example of a CA that qualified in the 2001 Census-based design but not in the 1991 design is Squamish, British Columbia. A new code was created for it, code 934.

Case C: Codes applicable in all reference years. Most codes have retained the same name and still apply because that area is still a CMA or CA on its own. The boundary may also have changed if there was a change in municipal jurisdictions; this is not indicated in the CMACA25 variable, but it is taken into account in the Census-specific mapping of CSDs to CMAs, CAs, and non-CMA-CA areas. The CMACA25 code 225 used to be named Sydney, Nova Scotia in the 1991 Census-based design, but it was renamed to Cape Breton due to amalgamation and municipal restructuring. This is an example of a name change combined with a boundary change. The description with this code indicates that it represented Sydney until 1998, and thereafter Cape Breton.

How does SLID do geography processing?

This section describes the input files and processing steps used by SLID for 2002 and also retroactively for 1999, 2000, and 2001 in the revised data.[9]

[9] For information on how SLID processed its geographical information using the 1991 Census-based design, please refer to “SLID Geography and its Impact on Low Income Measurement” by Cunningham, R., Lafrance, P., Rowland, J. and Murray, J., 1991, The Income and Labour Dynamics Working Paper Series, Statistics Canada, Catalogue no. 97-09.

Input files

There are four main input files required to process geographical information for each household. They are the Postal Code Conversion File (PCCF), the Geographic Attribute File (GAF), the Employment Insurance Region (EIR) file and the Telephone Billing file.

The **Postal Code Conversion File (PCCF)** produced by Geography Division contains the link between the six-character postal codes and Statistics Canada's standard census geographical areas (such as DAs, municipalities and CMA/CAs). It also provides the latitude and longitude coordinates for a point representing the approximate location of the center of the postal code area. In addition to being updated every five years to reflect new Census geography, the PCCF is updated every six months to take into account the postal code changes continually introduced by Canada Post Corporation (CPC).[10]

The **Geographic Attribute File (GAF)** contains the population and dwelling counts of the latest census, by Dissemination Area (DA). It also provides the correspondence or "link" between each DA and all higher geographic levels except the Employment Insurance Region. This includes information identifying urban and rural areas (urban/rural flag, urban area code and urban area population) and the longitude and latitude of a representative point in each DA.

The **Employment Insurance Region (EIR)** file provides the link between each DA and Employment Insurance Regions (EIRs), which are defined by Human Resources and Skills Development Canada (HRSD) and Social Development Canada (SD) (formerly Human Resources Development Canada). Eligibility for Employment Insurance benefits is based, in part, on the unemployment rate in the applicant's EIR.

The above three files (PCCF, GAF, and EIR) can be merged as one file, using the dissemination area as their common field. Whether the SLID data are linked to these files one at a time or after merging the three together does not really matter, for the purposes of simplifying the description of processing steps below, we can refer to them jointly as the PCCF-GAF-EIR file.

The SLID staff produces a single **Telephone Billing file** each year from the multiple telephone billing files received quarterly from approximately 22 companies that provide telephone services across Canada. Note that these telephone billing files do not contain the names of the subscribers nor any billing information. Only the telephone number, postal code and province fields are required for SLID processing.[11]

[10] The PCCF was first linked to the 1981 Census geographic areas and has undergone four "conversions" since then, to reflect the 1986, 1991, 1996 and 2001 censuses. For more information, please refer to "Postal Code Conversion File (PCCF), Reference Guide", Statistics Canada, Catalogue no. 95F0153GIE.

[11] Although useful, the telephone files have a few limitations which could prevent a link with survey respondents, or which might give an address other than the person's permanent residence. The bill may be sent to another residence or a post office, although this does not actually matter unless the postal code of that address is not the same as the person's residence. Depending on the telephone company, unlisted numbers may not be included, in which case a match would not be found.

Once these file have been obtained, the SLID geography processing for each household can proceed.

Steps to assign geography values

Step 1

After setting aside all records with no postal code in the current year (these will be processed in step 3 later), a match is done between households in the current year and households of the previous year by household ID, province and postal code. For all current-year households for which a match is found, census geographic codes are copied over from the previous year.

Before the unmatched records from this step are passed to the next step, they are checked for consistency of their postal code and telephone number with their province code. For example, households residing in Ontario must have a postal code that begins with a K, L, M, N or P, and their telephone area code must begin with 289, 416, 519, 613, 647, 705, 807, or 905. If either the postal code or the telephone number fails this check, then both are set to “missing” and only the province is retained. Either their geography variables will be assigned values in Step 3, or they will be imputed.

Step 2

Match the unmatched records from Step 1 with the PCCF-GAF-EIR file by province and postal code. All records with a postal code will find a match on the file, unless some invalid codes were not caught in Step 1. For simplicity, we can assume that all unmatched records at this stage have a postal code value of “missing”.

Step 3

Match the unmatched records from Step 2 and the records with no postal code (they were set aside in Step 1) with the Telephone Billing file, by their 10-digit telephone number. For successfully matched records, the postal code is added and then the record is re-matched with the PCCF-GAF-EIR file to obtain all the geographic values. Those that did not match at this stage are sent to Step 4.

Step 4

Those records that still do not have a postal code after Step 3 are compared to last year’s file at the same stage (i.e. just after Step 3) by household ID, address *excluding* postal code and phone number. If they are the same, then the geographic values are carried forward to the current year, to avoid changing the location of residence through re-imputation. The rest of the records will have their postal code value assigned by imputation.

Step 5

Impute the remaining unmatched records from Step 4. Please refer to the next section "Imputation methods". Once a postal code, CD and DA codes have been imputed, we match these records to the PCCF-GAF-EIR file to retrieve the rest of the geography values.

Geography processing is then complete except for some variables that are not directly provided by the PCCF-GAF-EIR file but which can be derived, such as the distance moved.

Table 4 gives the record counts for the processing steps above.

Table 4: Records assigned geography codes, by method (2001 Census-based design)

Reference Year	**1999**	**2000**	**2001**	**2002**
Total number of households in sample	35,761 (100%)	37,125 (100%)	36,444 (100%)	36,766 (100%)
Outside Canada records: All fields set to not applicable	305 (1%)	389 (1%)	481 (1%)	307 (1%)
Step 1: Codes copied from previous year, since the postal code and household ID are the same	*	25,491 (69%)	27,572 (76%)	14,646 (40%)
Step 2: Codes obtained by matching postal code to PCCF-GAF-EIR file, since postal code is already available and valid	34,064 (95%)	3,585 (10%)	6,084 (17%)	19,875 (54%)
Step 3: Codes obtained by using telephone number to get a postal code, then matched to PCCF-GAF-EIR file	800 (2%)	4,916 (13%)	728 (2%)	860 (2%)
Step 4: Codes copied from previous year's values since postal code missing but the address and household ID are the same	*	81 (0%)	180 (0%)	247 (1%)
Step 5: Postal code, CD and DA are imputed and then record is matched to PCCF-GAF-EIR file	592 (2%)	2,663 (7%)	1,399 (4%)	831 (2%)

*Not applicable because 1999 was the earliest year to be processed using the 2001 Census-based geography design.

Note: In step 1, the percentage of records having their geography codes copied from the previous year for 2002 is lower than for 2000 and 2001 because of the introduction of a new panel in 2002.

Imputation methods

The postal code imputation methods are presented below in the same order as they are used in SLID processing. These methods were used each year starting with processing of data for 2002, but they also apply to years 1999, 2000 and 2001 because the imputed geography data of those years were re-imputed in a historical revision. Prior to 2002 processing, SLID used the very simple method of randomly selecting any one of the postal codes that existed in the province.

The imputation methods described here are all within a given province. One main reason to restricting imputation to be within province was to respect the weighting procedure for the historical revisions of 1999 to 2001. Another reason for restricting imputation to be within a given province is because some records need their province to be imputed beforehand as the SLID staff was unable to trace them. The method used for province imputation is just to assume that the household still lives in the same province as the last known province.[12]

1. Longitudinal imputation – backward (POSTCI25=40)
    - This type of imputation can only be used when doing historical revisions since it takes known information from one year and applies it to one or more preceding years.
    - This method was used for the historically revised years, 1999 to 2001.
    - If a household record exists in the survey in at least two consecutive years and their address in the second year is known, but their address in the first of the two years is unknown, then it is assumed that the SLID staff lost contact with the respondent in the first year of missing data. This assumption is logical because often when the SLID staff cannot trace a respondent, it is because they moved. Another assumption is applied that they did not move between the two years, so the later year's address information is imputed to the previous year. [13]

2. Longitudinal imputation – forward (POSTCI25=43)
    - Only used for reference years 2000 and on – after re-processing 1999 data.
    - If the household needs their postal code to be imputed two years consecutively, then they are given the same postal code for these two years. This is because no information is known whether the household did or did not move. The SLID staff did not want to re-impute a postal code for them and re-locate the household. This necessarily assumes that they did not move.

3. Imputation using six-digit telephone number (area code + exchange bank) (POSTCI25=41)
    - For those with a telephone number in the current reference year, extract the first six digits (area code + exchange bank). For example, the first six digits of a telephone number for someone residing in Ottawa might be (613744), where (613) is the area code for Ottawa and (744) is the exchange bank. We match these with the six-digit telephone numbers on the telephone billing file which,

[12] The imputation of the household location within province of the last known address is an artificial limitation since in reality it is possible that respondents move to another province. Ideally, inter-provincial movement would be allowed for in the imputation method. But since it requires a lot of research into different methods before it can be implemented, it was excluded from the recent changes in geography processing, which already included a new Census design, improved imputation methods, and historical revisions. However, the intention is to look at it in the near future.

[13] Postal codes were not affected by the switch in geography design.

was first merged with the PCCF. If there are many postal codes for that area code and exchange bank, we randomly choose one of them.

- If in the current reference year the telephone number is unidentifiable or there is no telephone number for the respondent, then method 4 is used.

4. Distance imputation using previous year's postal code (POSTCI25=42)
    - This method is used for records that were in the survey the previous year and did not have to be imputed that year; therefore, one knows with certainty what their location was one year ago. The first stage to this method is to impute how far they have moved. An assumption is made that they had moved because the SLID staff was not able to trace them for the survey. After this first stage, one knows all the possible longitude and latitude coordinates they could have moved to, which would be consistent with the distance moved. The second stage is to select one of these pairs of coordinates and then use it to obtain the postal code. Below are a few more details on how the distance imputation is done.
    - For each SLID respondent that was in the survey in both the current and preceding years with a valid postal code in both those years (i.e. without imputation) and whose postal code changed, their "distance moved" is calculated. All these values are sorted by province and move type (i.e. urban-urban, urban-rural, rural-urban or rural-rural). Then, for each household requiring imputation, randomly select one of the values of "distance moved" from all those with the same province and move type as the household. This will be the donor used to establish approximately how far the household requiring imputation has moved.
    - The PCCF is used to obtain the list of longitude and latitude coordinates of all the postal code areas within the province; it gives just one pair of coordinates per postal code area. Then, for each household requiring imputation, find all the possible destinations that they could have moved to within the same province given their distance moved and previous location. In order to get a number of destinations, the distance moved is extended to be within, for example, 5 kilometres plus/minus around the donor "distance moved". One of these destinations is randomly chosen to be the destination that the household actually moved to. The postal code can then be obtained from the coordinates.

5. Random imputation of a postal code/DA/CD (POSTCI25=44)
    - This last resort method is applied only if no postal code can be found from any of the above methods. A postal code within province is randomly chosen.
    - Records that are left for random imputation are usually the ones with no postal code or telephone number. For many records, they only contained their household ID and their province of residence while all the other fields of information are empty.

The variable POSTCI25 serves to record how each household record obtained its final postal code value. The various categories are shown in Table 5 below. The possible categories include the different types of derivation or imputation that were applied, depending on the year. Codes 30, 31 and 32 now apply only to reference years 1992 to 1998, which are still based on the 1991 Census geography in SLID. Codes 40 to 44 make reference to the 2001 Census geography and SLID's updated imputation methods. Code 40 applies to only 1999 to 2001 and code 43 only applies to 2000 onwards. Also, codes 41, 42 and 44 apply to the years starting with 1999.

Table 5: Categories of variable POSTCI25: How postal code was assigned

Code	Description
10	Collected postal code
20	Postal code derived from telephone number
30	Federal electoral district/Enumeration area assigned by Geography Division using address (RY1992-RY1998)
31	Federal electoral district/Enumeration area imputed by Geography Division using province (RY1992-RY1998)
32	Federal electoral district/Enumeration area assigned from Stratum-Type-Cluster of LFS Frame (RY1992-RY1998)
40	Longitudinal imputation – backward (RY1999-RY2001)
41	Imputation using reference year 6 digit telephone number (area code + exchange bank) (i.e. 613-951) (from RY1999 on)
42	Distance imputation using previous year's postal code (from RY1999 on)
43	Longitudinal imputation – forward (from RY2000 on)
44	Random imputation of postal code/DA/CD (from RY1999 on)

Note: RY stands for reference year. The reference years indicated in brackets implies it is used only for those reference years.

Table 6 below contains the number of records with postal codes imputed by the different methods.

Table 6: Counts by imputation methods (POSTCI25)

Code	**Imputation methods**	**1999**	**2000**	**2001**	**2002**
40	Longitudinal imputation – backward (RY1999-RY2001)	41 (6.9%)	1,031 (38.7%)	40 (2.9%)	…
41	Imputation using reference year 6 digit telephone number (i.e. 613-951) (from RY1999 on)	187 (31.6%)	793 (29.8%)	146 (10.4%)	179 (21.5%)
42	Distance imputation using previous year's postal code (from RY1999 on)	303 (51.2%)	751 (28.2%)	982 (70.2%)	483 (58.1%)
43	Longitudinal imputation – forward (from RY2000 on)	…	0 (0.0%)	151 (10.8%)	20 (2.4%)
44	Random imputation of a POSTCD/DA/CD (from RY1999 on)	61 (10.3%)	88 (3.3%)	80 (5.7%)	149 (17.9%)
…	Total	592 (100.0%)	2,663 (100.0%)	1,399 (100.0%)	831 (100.0%)

Note: RY stands for reference year. The reference years indicated in brackets implies it is used only for those reference years.

While introducing these improved methods for replacing missing values of geographic information, the SLID staff noticed an additional improvement that could be made.

The distance imputation method described above for 1999 to 2002 used an approximate formula:

(1) $DM=75*SQRT[(LATRES_{RY} – LATRES_{RY\text{-}1})^2 + (LNGRES_{RY} – LNGRES_{RY\text{-}1})^2]$,

where DM=distance moved
SQRT=square root
LATRES=latitude of residence
LNGRES=longitude of residence
RY=current reference year
RY-1=previous reference year

This formula (1) tends to underestimate the distance moved by approximately 33% for low latitudes and overestimate the distance moved by approximately 20% for high latitudes.

Starting in 2003, it was replaced by a more precise one:

$$\begin{aligned}(2)\ DM=111.32^*\ \arccos\ [&\sin(LATRES_{RY})^*\ \sin(LATRES_{RY\text{-}1})+\cos(LATRES_{RY})^* \\ &\cos(LATRES_{RY\text{-}1})^*\ \cos(LNGRES_{RY} – LNGRES_{RY\text{-}1})]\end{aligned}$$

This formula (2) calculates the distance along a “great circle arc”, i.e., the shortest surface distance between any two points on a sphere (i.e. assuming the earth to be a perfect sphere).

Impact on data due to the switch to 2001 Census-based design

The data prior to 1999 is based on the 1991 Census-based design. Since the data for reference years 1999 and later is based on the 2001 Census-based design, there is a structural break in the data series between 1998 and 1999. Since many geographical boundaries have changed, as well as many name and code changes have occurred, the geography codes for some households between 1998 and 1999 may be different even if the household did not move.

Furthermore, due to the change in imputation methods, some households that were imputed in reference years 1999, 2000 or 2001 may be located in a different location when examining the data for both the 1991 and 2001 Census-based design.

Updating geographical information can impact low income rates in particular. For example, the family size and the community size are used to find the appropriate cutoff. Then the family income is compared to that cutoff. If a family low-income rate is being calculated, then the family is counted as being in low income if its income is less than the cutoff. If a person low-income is being calculated, then all persons in the family are counted as being in low income if the family income is less than the cutoff.

Low income cuttoffs (LICOs) are established using data from the Family Expenditure Survey, now known as the Survey of Household Spending. LICOs convey the income level at which a family may be in straitened circumstances because it has to spend a greater proportion of its income on necessities than the average family of similar size. Specifically, the threshold is defined as the income below which a family is likely to spend 20 percentage points more than the average family. There are separate cutoffs for seven sizes of family – from unattached individuals to families of seven or more persons – and for five community sizes – from rural areas to urban areas with a population of more than 500,000.

Table 7 below contains low-income rates (the proportion of persons living in low income) for reference years 1999, 2000 and 2001 for both the 1991 and 2001 Census-based design. This table indicates that the low-income rates are not significantly affected by the change from using the 1991 to the 2001 Census-based design to process the geography variables.

Table 7: Low-income rates for all persons in 15 major CMAs in Canada

	1999			
	Before-tax		**After-tax**	
Census-based design	**1991 Census**	**2001 Census**	**1991 Census**	**2001 Census**
Montreal	23.9%	24.1%	19.3%	19.4%
Ottawa-Gatineau	17.7%	17.9%	14.6%	14.7%
Winnipeg	19.2%	19.3%	15.7%	15.9%
Halifax	17.1%	17.1%	13.3%	13.5%
Quebec City	18.2%	18.6%	14.0%	14.4%
Toronto	14.8%	14.8%	11.2%	11.2%
Hamilton	12.1%	12.2%	9.2%	9.2%
St. Catherines-Niagara	9.6%	9.5%	5.5%	5.5%
Kitchener	10.8%	10.8%	8.0%	8.0%
London	16.0%	16.1%	12.8%	12.8%
Windsor	12.8%	13.4%	8.1%	8.3%
Calgary	13.5%	13.5%	11.2%	11.3%
Edmonton	15.8%	15.8%	12.5%	12.5%
Vancouver	20.2%	20.2%	16.5%	16.6%
Victoria	15.4%	15.2%	10.6%	10.6%
	2000			
	Before-tax		**After-tax**	
Census-based design	**1991 Census**	**2001 Census**	**1991 Census**	**2001 Census**
Montreal	21.2%	21.7%	17.6%	17.6%
Ottawa-Gatineau	14.4%	14.8%	12.5%	12.7%
Winnipeg	18.3%	18.8%	14.8%	15.3%
Halifax	14.7%	15.0%	12.0%	12.2%
Quebec City	16.7%	17.3%	13.8%	14.1%
Toronto	13.3%	13.4%	10.5%	10.7%
Hamilton	11.5%	11.5%	9.0%	9.1%
St. Catherines-Niagara	7.5%	7.4%	5.5%	5.5%
Kitchener	10.0%	10.0%	8.1%	8.1%
London	12.0 %	12.0%	9.8%	9.4%
Windsor	11.6%	11.4%	7.8%	7.8%

Calgary	12.2%	12.2%	8.8%	9.0%
Edmonton	17.0%	17.0%	12.2%	12.2%
Vancouver	17.3%	17.3%	14.4%	14.5%
Victoria	18.3%	18.3%	14.8%	13.9%
	2001			
	Before-tax		**After-tax**	
Census-based design	**1991 Census**	**2001 Census**	**1991 Census**	**2001 Census**
Montreal	19.0%	19.0%	14.8%	14.9%
Ottawa-Gatineau	11.4%	11.6%	8.1%	8.3%
Winnipeg	16.1%	16.5%	11.2%	11.6%
Halifax	14.2%	14.4%	11.7%	11.7%
Quebec City	17.2%	17.7%	13.1%	13.6%
Toronto	11.0%	11.1%	8.4%	8.4%
Hamilton	10.6%	10.6%	6.9%	6.9%
St. Catherines-Niagara	7.4%	7.4%	4.8%	4.5%
Kitchener	8.1%	8.1%	4.5%	4.5%
London	11.6%	11.6%	7.9%	7.9%
Windsor	9.4%	9.6%	7.9%	7.9%
Calgary	11.7%	11.7%	8.9%	9.0%
Edmonton	15.1%	15.2%	9.8%	9.8%
Vancouver	17.4%	17.4%	13.3%	13.3%
Victoria	14.5%	14.5%	10.1%	10.1%

Conclusion

The SLID staff knew at some point that it would be necessary to align the SLID geography definitions to that of a more recent Census. To that end, the SLID geography definitions were changed from 1999 onward, while overlapping the definitions for 1999, 2000 and 2001. The impact, as seen in the previous section, is modest and will allow the SLID staff more breathing room until the next change, perhaps following the 2006 or 2011 Census.

Though imputation was emphasised in this document, the grand majority of geographic coding is done by copying codes from a previous year or by matching a respondent’s valid postal code to geography files. The rest are handled by fairly sophisticated methods which ensure that the user receives the best geographic coding that the SLID staff can give them.

References

Census Operations Division. 1992. “1991 Census Dictionary, Reference”. Statistics Canada, Catalogue No. 92-301E.

Census Operations Division. 1999. “1996 Census Dictionary, Final Edition Reference”. Statistics Canada, Catalogue No. 92-351-UPE.

Census Operations Division. 2002. “2001 Census Dictionary, Reference”. Statistics Canada, Catalogue No. 92-378-XPE.

Cunningham, R., Lafrance, P., Rowland, J. and Murray, J. 1997. “Slid Geography and its Impact on Low Income Measurement.” The Income and Labour Dynamics Working Paper Series, Statistics Canada, Catalogue No. 97-09.

Geography Division. 1995. “Postal Code Conversion File (PCCF), User Guide.” Statistics Canada.

Geography Division. 2003. “Postal Code Conversion File (PCCF), Reference Guide.” Statistics Canada, Catalogue No. 92F0153GIE.

Income Statistics Division. 2002. “Income in Canada, 2002.” Statistics Canada, Catalogue No. 75-202-XIE.

Services, Publications Division. 1992. “Standard Geographical Classification SGC 1991, Volume 1, The Classification.” Statistics Canada, Catalogue No. 12-571.

Standards Division. 2002. “Standard Geographical Classification SGC 2001, The Classification.” Statistics Canada, Catalogue No. 12-571-XPB.