Catalogue no. 91F0015M — No. 12
ISSN 1205-996X
ISBN 978-0-660-06596-0

Demographic Documents

Using family-related variables from the Census of Population and the National Household Survey microdata files



by Heather Lathe, Anne Milan and Nadine Laflamme

Release date: December 22, 2016

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

Standard table symbols

The following 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
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2016

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement.

An HTML version is also available.

Cette publication est aussi disponible en français.

Using family-related variables from the Census of Population and the National Household Survey microdata files

by Heather Lathe, Anne Milan and Nadine Laflamme

Abstract

Family-related variables are a significant part of the Census of Population, but in order to use them appropriately for research purposes, it is important to understand them. This article provides information on using family-related variables from the microdata files of the 2011 Census and earlier censuses, as well as those of the 2011 National Household Survey (NHS). These microdata files vary in their attributes depending on whether they are located internally at Statistics Canada, in the Research Data Centres (RDCs), or whether they are public-use microdata files (PUMFs). This article compares these three versions of the microdata files including their similarities and differences. It explains technical aspects of using family-related variables such as how additional family variables (using the concepts of census families or economic families) can be created for analytical purposes, including the creation of multi-level variables. This article is a useful supplement to technical documentation already with the 2011 Census microdata files and those of previous censuses.

Introduction

The Census of Population, as well as the National Household Survey (NHS) which was carried out in 2011, provides a statistical portrait of the population and is a key source of detailed data for small groups as well as many levels of geography. These data can be used for planning and decision-making purposes by government, business, media, academics and the general public as well as others interested in social, demographic, and economic information.

Family-related variables are an important component of the census and NHS data. They can also be complex to understand and use in analysis. In conjunction with other reference materials,[1] this article provides a source of information for researchers interested in conducting family-related research with the microdata files.

Section 1 begins with an introduction to multi-level analysis before providing an overview of family concepts.

Section 2 outlines the similarities and differences between the Census Program's Dissemination database (available to Statistics Canada employees only), the Research Data Centre (RDC) microdata files and the public-use microdata files (PUMFs).[2] Researchers can use the public-use files for exploratory analysis before submitting a proposal to use the RDC microdata files. The hierarchical or 'multi-level' nature of the files is discussed in this section due to its importance for family variables.

Section 3 provides an overview of the family-related variables in the 2011 Census and NHS, highlighting those that are most central to the concepts of census families and economic families. This section concludes with a description of the identifier variables, given that identifiers are important when using microdata files for family analysis.

Section 4 explains how additional family variables can be created by analysts for their research, particularly multi-level variables. Multi-level variables are those that cross units of analysis, for example, when a higher-level characteristic such as household income is applied to a person or individual record.[3] The creation of additional variables is possible due to the hierarchical content of the census and NHS microdata files.

Section 5 contains general technical aspects of the census and NHS databases: the selection of the appropriate population or universe for analysis; the application of weights; and the use of identifier variables.

1. The *Census Dictionary* of each census and the *Families Reference Guide* of 2006 and 2011 cover details pertaining to historical comparability, conceptual changes and data quality of a given reference year.
2. Public-use microdata files are also referred to as PUMFs or public-use files.
3. The terms "person" and "individual" are used interchangeably when referring to the unit of measurement found in a data file. However, in most cases, the term "person" will be used except when referring to the Individuals PUMF.

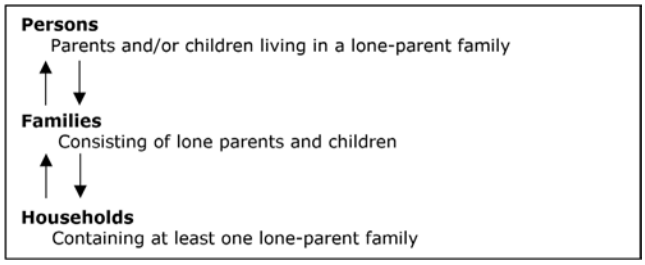
1. Concepts for family analysis

1.1 Multi-level analysis

The **dwelling** is the collection unit for the census and NHS. All persons living in a dwelling as their usual place of residence make up the **household** of that dwelling. **Families** are then identified among the household members using Question 6 on the census questionnaire, 'Relationship to Person 1', which asks for the relationship of each member with respect to a single reference person in the household. This information is used together with the responses of each person to sex, date of birth and marital status to derive family variables.

Even if a given research topic is broadly related to families, the unit of analysis can be persons, families or households. The choice to use persons, families, households or a combination thereof depends on the particular research question.

Figure 1
Multi-level analysis possibilities for lone-parent families



As an example, the topic of lone-parent families can be analyzed at the level of persons, families or households (Figure 1). In the first case, in order to count lone parents, data would be examined at the person level. In the second case, in order to examine lone-parent families, data at the family level is appropriate, given that family variables have already been created in the database to relate different household members to each other in a family concept. In the third case, where the data of interest are lone-parent family households, information at the family level must be 'brought up' to the household level.

Note that it is not necessary to restrict the family-level analysis to families, to the exclusion of persons not in a family. For example, family size can be 1 for persons not in a family. 'Family income' can be equal to the individual's income if the person is not in a family.

1.2 Census family and economic family concepts

In the standard Statistics Canada definitions for families, two complementary concepts exist: **census family** and **economic family**. Most analysis by Statistics Canada on family characteristics using census or NHS data will be based on at least one of these two concepts.

The census family is the narrower concept. It corresponds to the concept of a family nucleus that the United Nations (2015) recommends as the primary aspect of household composition to be considered for census purposes. It is defined as either a couple with or without children or a lone parent living with his or her children, where each child is not living with his or her own married spouse or common-law partner or child.

The economic family encompasses any two or more people living together who are related to each other by blood, marriage, common-law union, adoption or a foster relationship. The economic family concept of 'child' generally includes a greater number of older sons and daughters compared with census families given that children in an economic family can themselves be members of couples or lone parents, even while living with one or both parents.[4]

Both the definitions of who makes up a census or economic family and the role or status of each person in the family are important for understanding which concept to use for a particular research question. Additionally, research needs will determine whether persons not in census or economic families should be included as part of the study population.

4. Although there is no age restriction on children in the generic concepts of census families and economic families at Statistics Canada, researchers frequently add an age restriction in order to focus the analysis on families with children below a given age (for example, families with at least one child under age 2, 6, 18 or 25). There are variables in the census and NHS that identify whether a census family contains at least one child that corresponds to these age cut-offs.

As an example of the analytical difference, census family concepts could be used for studying family incomes, in order to distinguish between the family resources of the census family and those of the broader economic family to which the census family belongs, in situations where a census family is living with additional relatives. The census family concept is similar to the concept of the family for tax purposes, especially when an age restriction is applied to children, such as ages 0 to 17.

The economic family concept, however, is often used for analysis pertaining to incomes, given the assumption that all related persons living in the same dwelling share many financial and material resources. In this case, 'persons not in economic families' may be counted as economic units along with economic families, in order to cover the income portrait of the entire population.

The concept of a household is also useful for studying families and living arrangements. In the example of income-related topics, the economic situation of one-person households could be of analytical interest in comparison with other household types.

The hierarchy of households, economic families, census families and individuals is shown in Figure 18 of the *2011 Census Dictionary*, http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/figures/figure18-dict-eng.cfm. It can also be expressed in several 'one-to-many' statements:

- 1 census family: 2 or more persons.
- 1 economic family: 2 or more persons.
- 1 economic family: 0, 1 or more census families. Persons not in a census family can also be present.
- 1 household: 1 or more persons.
- 1 household: 0, 1 or more census families. Persons not in a census family can also be present.
- 1 household: 0, 1 or more economic families. Persons not in an economic family can also be present.

Box 1

Family data changes over time—A quick reference

The concepts of census family, census family status and census family structure have remained the same in the census since 2001. Prior to 2001, the census family concepts were the same from 1976 to 1996. The changes made to the census family concepts in the 2001 Census are described under 'Census family' in the *2011 Census Dictionary*, http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/fam004-eng.cfm and under 'Historical comparability' in the *Families Reference Guide*, 2011 Census, http://www12.statcan.gc.ca/census-recensement/2011/ref/guides/98-312-x/98-312-x2011005-eng.cfm#a5.

More information about family concepts over time is contained in the Appendix of the article, 'Enduring Diversity: Living Arrangements of Children in Canada over 100 Years of the Census', no. 11, *Demographic Documents* (Statistics Canada Catalogue no. 91F0015M), http://www.statcan.gc.ca/pub/91f0015m/91f0015m2014011-eng.htm.

Family characteristic	First year of data availability
Same-sex common-law partners or couples	2001
Same-sex married spouses or couples	2006
Foster child as member of an economic family	2011
Stepfamilies, intact families, and individuals in these families	2011

2. Comparisons between Statistics Canada's Dissemination database, the Research Data Centre microdata files and the public-use microdata files

2.1 Complete microdata files

The Census Program's **Dissemination database**, which is internal to Statistics Canada, contains microdata files of all short-form and long-form records—referred to as 100% data—and sample data records only, for the nine censuses from 1971 to 2011. The sample data carry both short-form and long-form variables, while the 100% data carry only short-form variables. The sample data for 2011 are the NHS records combined with their census characteristics.

Over time, **Research Data Centre (RDC) microdata files** have been created for the same years, to meet the needs of researchers outside Statistics Canada. The RDC files contain the same records and content as the Dissemination database for sample data.[5]

While access to Statistics Canada's Dissemination database is restricted to analysts within the department, access to the RDC files is achieved through an on-line proposal process.[6]

One technical difference exists between the RDC files and the long-form Dissemination database. While it does affect how the data need to be extracted to some extent, it does not affect the types of analysis that can be done. The Dissemination database stores variables on five separate files or 'tables' for each unit of analysis: dwelling, household, person, census family and economic family. Identifier variables or 'keys' allow the units of one file, such as persons, to be linked with the corresponding units of another file, such as families. Consequently, the database is said to be 'relational' in structure.

In contrast, in the RDC file, the different tables have been merged into a single flat file containing only person records. The characteristics of the family are attached to the records for each person of that family, and similarly for household characteristics. The identifier variables for families and households have not been removed from each person record, and this is why it is still said to be 'hierarchical' in terms of its content.[7]

One advantage for data users of having the RDC file and Hierarchical PUMF as person-level files is that all the variables are together, so some family analysis can be carried out without merging the separate files. It is possible to either use all the family and household variables at the person level as is, or to select one person per family or one person per household to use them at the family or household level. This assumes, however, that the analyst does not need to derive any new variables linking person and family characteristics together.

The disadvantage of a person-level file is that in order to interpret the data correctly, data users must know whether a particular variable represents a person-level characteristic or a family or household characteristic attached to each person record. The variable descriptions indicate to which unit of analysis the characteristic applies.

Due to the differences between the National Household Survey and the 2011 Census (a short-form census that year), the 2011 Census RDC file, which contains only short-form characteristics, was made available in the RDCs in November 2014. It was provided as a one-fifth sample to reduce its size (with a weight variable to compensate). For simplicity, the 2011 Census RDC file is not mentioned further in this article. Apart from the fact that it is a sample, all details about it are the same as for the 2011 Census Dissemination database.

5. With the exception of data for unoccupied dwellings, which are only in the Dissemination database. Data for households are equivalent to data for occupied dwellings.
6. See The Research Data Centres (RDC) Program (www.statcan.gc.ca/eng/rdc/index) for more information.
7. For more information on hierarchical files that have been 'rectangularized' as flat files, refer to Peller (2012).

2.2 Public-use microdata files

Public-use microdata files (PUMFs) are available for the 2011 NHS and the long-form censuses back to 1971. Public-use files are more restricted than the complete files in both content and size. They consist of a relatively small sample of records from the original files (with large weights, to compensate), and they have less geographic detail and fewer variables on other characteristics as well.

Furthermore, the variable categories have been reduced (for example, age is only available as age groups) or the microdata have been modified to ensure that individual respondents cannot be identified. Access to PUMFs is provided via the Data Liberation Initiative of participating post-secondary institutions.[8] Depending on a given research project, external analysts may find that using PUMFs is sufficient, or they can use them for exploratory analysis before submitting a proposal to use the RDC microdata files.

Prior to 2006, all public-use files existed as separate household, family and person-level files, and there was no means to link records between them. This limited the types of analysis that could be conducted. However, since 2006 there is, instead, a hierarchical file and an individuals file.

The **Hierarchical Public-Use Microdata File** is structured the same way as the RDC file, in that all variables pertaining to family- and household-level characteristics have been attached to the person records. It has only a small number of variables for household and family characteristics (and person-level characteristics), but it still has the family and household identifiers.

In contrast, the **Individuals PUMF** has more variables than the Hierarchical PUMF. This makes it more suitable for analysis at the person level when household and family characteristics are not required. It also has a larger size, containing about 2.7% of the original long-form household records, compared to 1% for the Hierarchical PUMF. However, no additional multi-level information can be obtained from the Individuals PUMF, given that the identifiers that would allow links between individuals are not provided.

8. See the Data Liberation Initiative (DLI) (www.statcan.gc.ca/eng/dli/dli) for more information. See also The Research Data Centres (RDC) Program (www.statcan.gc.ca/eng/rdc/index) for more information about the differences between non-confidential public-use files and the confidential RDC files.

3. Variables for family analysis in the census and NHS

Section 3.1 describes the main family-related variables and includes some points about the basic processing applied to the data, which may be helpful for programming new variables. Section 3.2 explains the identifier variables, which are important for combining the different units of analysis.[9]

3.1 Family-related variables and processing

A list of all the demographic and family variables for 2011 is found in the Appendix. The variables are presented in order of the unit of analysis to which they apply: persons, families and households. Only a few of the variables represent direct responses from the questionnaire. Most of them are derived from a combination of the responses of each person to the demographic questions (age, sex and marital status) and the relationship question for all members of the household.[10]

Almost all of the analysis and tables produced by Statistics Canada for the official 2011 Census release on families, households and marital status[11] can be produced using this collection of variables. Researchers may also derive their own variables to meet their analytical needs (information about deriving variables is provided in Section 4).

The basic **person-level** variable on census family characteristics is Census family status or CFAMST. The variable Household living arrangements (CFSTAT) indicates whether each person not in a census family is living with other relatives, non-relatives only, or alone. Economic family status, or EFAMST, is the equivalent of CFAMST, but for the economic family.

The basic **family-level** variable on census families is Census family structure or CFSTRUCT. Economic family structure (EFSTRUCT) is the equivalent variable for the economic family concept.

The five variables CFAMST, EFAMST, CFSTAT, CFSTRUCT and EFSTRUCT all have large category sets because the basic family concepts that they represent have been crossed by sex and marital status to obtain details such as whether couples are married or common-law, whether they are opposite-sex or same-sex, and whether lone parents are male or female. For 2011, new simplified versions of these variables were added to the database (and RDC files) to represent only the basic family concepts without any extra details. The variable names are the same but end in 'SIMPLE', such as CFAMSTSIMPLE.

The basic **household-level** variable on census families is Household type (HHTYPE). It divides households into family households, which are composed of at least one census family, and non-family households. In addition, it contains some information on census family structure and whether persons not in a census family are present. HHTYPE is only found in the Individuals PUMF; however, it can be derived in the other files (see Example 4.3.2 in Section 4.3).

Marital status (MARSTH) is the main marital status variable and it is derived from Legal marital status (MARST) as well as Common-law status (COMLAW). The variable HWCLPR indicates when a person is married but their spouse is not a usual resident of the same household (for reasons which exclude marital separation, such as illness, work or school). This allowance in the marital status variables that a married spouse can be absent differs from the family variables. In the family variables, a 'couple' requires that both individuals be present in the household. Consequently, the counts of married spouses will differ when using family variables such as CFAMST/CFSTAT compared with MARST/MARSTH. (They may also differ if published for a different universe, as explained in Section 5.1.)

9. In this article, all variable names are shown in upper-case letters for simplicity.
10. In the census and NHS, data on individuals are paired in a 'family formation' process based on probability rules, whereby two married spouses or two common-law partners are paired to form a couple and the couples or lone parents are paired with their children to form census families with children. The order of reporting of household members on the questionnaire is sometimes taken into account to help resolve complex or unclear situations. The census family variables are then derived, and they are used in the derivation of economic family variables.
11. Refers to release date of September 19, 2012.

The variable R2P1, which comes from the question on 'Relationship to Person 1' of the household, shows all the types of relationships after data capture and coding of the 'other-specify' response category in the question. It only appears in the RDC files as the result of its inclusion in the Dissemination database, where it was required for various technical purposes. R2P1 is not recommended for use for several reasons, notably, some response categories have been processed for dissemination only as grouped values and, most importantly, because this variable does not reflect the full set of edits that are applied during the family derivation process.

Family characteristics for persons living in seniors' residences, a type of non-institutional collective, are available in the 100% Dissemination database of 2011 and the 2011 Census RDC file which was mentioned in Section 2. Refer to that RDC code book for more information.

Certain edits have been applied to the data to ensure that they appear reasonable for all members of a household. For this purpose, an 'adult' is defined in the census as being 15 years or older.

- There is a minimum age difference of 15 years permitted between a parent and a child and a maximum age difference of 50 years between a mother and a child. However, starting in 2011, these edits are not necessarily applied to both parents in a couple as a result of the introduction of new son-daughter categories which permit the identification of stepchildren. Also, in the current processing environment, fewer edits of this type are applied to private households containing nine or more people than to households containing eight people or less.
- All private households have at least one adult.
- Age 15 is the minimum age requirement for a person to be married or living common law or to have his or her own child in the household.
- A census family always has at least one adult, which follows conceptually from the previous point.
- An economic family always has at least one adult, and the reference person of the economic family is always an adult.

3.2 Identifier variables

Persons in the same household, persons in the same family, or families in the same household can be determined by using the identifier variables provided in the file. This applies regardless of whether the file is relational in structure (the Dissemination database) or combined in a single flat file (the RDC and Hierarchical PUMF files), although the way the identifiers are used in programming is slightly different.

The household identifier may be called ID, HH_ID, or FRAME_ID, depending on the file. The identifiers for the person, census family and economic family are PP_ID, CF_ID and EF_ID, respectively. They are each unique across all the records for that type of unit: each household in Canada has a different value of HH_ID, etc. Furthermore, each person not in a census family has a unique value of CF_ID and each person not in an economic family has a unique value of EF_ID, so that they can be included as family units of one person (if relevant for the analysis), but also so that they don't appear to belong to one family.

The variables PP_ID, CF_ID and EF_ID do not exist on the census RDC files prior to 2011 or the Dissemination database prior to 2006. Data users must derive equivalent variables using the identifiers that are available for those years, which are PERSNO, C_FAM and E_FAM.[12] These variables are unique only within a household. To make persons or families unique within the entire file, it is necessary to precede their identifier value with that of the household to which they belong. This can be done either by concatenating the variables or by using the following formulations.

12. PERSNO, C_FAM and E_FAM exist in the RDC and Dissemination database files for every census since 1981. C_FAM and E_FAM may be found with the extensions 'PP', 'CF' or 'EF' but the values are the same. The identifiers PPNUM, HHNUM, CFNUM and EFNUM may still exist on the files of recent censuses, and they are valid for use with the geographic identifier PRCDDA, but they are not necessary in the presence of the other identifiers.

- The new identifier for persons (let's call it PP_ID) is created by multiplying HH_ID by 10,000 in order to add four digits to make space for PERSNO: PP_ID = HH_ID * 10,000 + PERSNO.
- The new identifier for families can be created in the same manner. It is shown here for census families (and called CF_ID), but economic families would be the same. It should also uniquely identify persons not in families.

  CF_ID = HH_ID * 10,000 + C_FAM, where C_FAM has a value between 1 and 99 for a family; and
  CF_ID = HH_ID * 10,000 + C_FAM + PERSNO, where C_FAM has a value of 0 for a person not in a family, which is the same as: CF_ID = HH_ID * 10,000 + PERSNO.

  In fact, the formula that was used for CF_ID and EF_ID in the more recent microdata files is three digits longer in order to include persons in collectives, for whom C_FAM and E_FAM both equal 999 (see Section 5.1 for more information about collective dwellings):

  CF_ID = HH_ID * 10,000,000 + C_FAM * 10,000, where C_FAM is between 1 and 99;
  CF_ID = HH_ID * 10,000,000 + C_FAM * 10,000 + PERSNO, where C_FAM = 0 or 999.

4. Examples of multi-level analysis

4.1 Examples of single-level data extractions

The first example below shows how family characteristics are extracted using person-level variables and counting persons as the unit of analysis. It can be carried out on the person table of the Dissemination database, or it can be carried out using the RDC file or the Hierarchical PUMF, given that both of these files have all the variables stored at the person level.

Example 4.1.1: single-level data extraction at the person level

Objective: The number of children under age 15 living with two parents.

This is the same number as the population under age 15 living in two-parent census families (including grandchildren living without their parents but with two grandparents). The variable CFSTATSIMPLE for Census family status, which is a person-level characteristic, contains a category for this data extraction: 'child of a couple' in a census family. It can be crossed by AGE (or AGEGR5) to apply the age criterion. (Note, however, that in order to exclude the cases where the two parents are grandparents, it would also be necessary to cross by CFAMST and exclude those children with CFAMST='Grandchild in a census family with no parent of grandchild present'. Variable CFAMST is not available in the Individuals PUMF or Hierarchical PUMF of 2011.)

The second example is also a single-level data extraction, but at the family level. It can be carried out on the family table of the Statistics Canada Dissemination database. It can also be carried out on the RDC file, but on that file it is necessary to first select one person per family to serve as a family record. The selection criteria are provided in Section 4.3.

Example 4.1.2: single-level extraction at the family level

Objective: The number of couples where both persons in the couple are aged 65 or older.

First of all, each couple corresponds to a couple census family, with or without children. Couple families can be identified with CFSTRUCTSIMPLE. The age restriction of each person in the couple can be applied by crossing this variable with CFAGE1STPRSN and CFAGE2NDPRSN, which indicate the 5-year age group of the first person and the second person, respectively. All three variables are at the family level, so this is still a single-level extraction, even though there are characteristics from the person level (age group) and the family level (census family structure). If the variables CFAGE1STPRSN and CFAGE2NDPRSN do not already exist, which is the case for the Hierarchical PUMF, then a multi-level extraction is needed to cross CFSTRUCTSIMPLE with AGE or AGEGR5 (not shown).

4.2 Examples of multi-level data extractions

Examples 4.2.1 and 4.2.2 provide illustrations of multi-level extractions. These examples assume that the variables pertaining to persons, households and families are contained on separate files, as in the Dissemination database. In the case of the RDC file and the Hierarchical PUMF, the analyst can first create a family-level file and a household-level file by selecting one person per family or household, using the criteria stated in Section 4.3. An illustration of creating these higher-level files is also provided in Section 5.3 (Example 5.3.1). In addition, any input variables shown in these examples but not provided in the Hierarchical PUMF have to be derived first.

Example 4.2.1: multi-level extraction at the family level

Objective: The number of female lone-parent families crossed by whether the mother in these families is in the paid labour force (employed or unemployed), the age of the mother (by five-year age groups) and whether she has any children under the age of six.

Female lone-parent families can be identified using CFSTRUCT. Variable CFAGE1STPRSN can be used to obtain the (five-year) age group of the mother. The age of the youngest child living with her, or specifically, whether that child is less than six, can be known using the relevant categories of variable CFKIDAGEMINGR. Then the labour market characteristics, as represented by person-level variables under the labour topic, need to be attached to each mother in the lone-parent family. This is done by carrying out a merge between the family-level table and the person-level table. The mother in the family is identified at the person level with variable CFAMST, selecting on CFAMST=Female lone parent. Her labour market characteristics are crossed with this value of CFAMST at the person level. Then they are attached to each family by carrying out a merge between persons and families based on the census family identifier variable CF_ID. The merge should retain the record for each family, while dropping the record for each person.

Example 4.2.2: a multi-level extraction at the household level

Objective: The creation of a variable on household type with values that show: a) whether the household has only one census family and the basic structure of that family and whether additional persons are present in the household; b) whether it is a multi-family household; and c) whether it has no census family and in that case, whether it is a household of one person or two or more persons.

The household will be the unit of analysis for this variable. A multi-level merge, using the household identifier HH_ID (or ID, depending on the file) is required to merge the household and family levels. Only households should be retained as records in the resulting output file. Certain family characteristics have to be retained (CFSTRUCTSIMPLE for census family structure and CFCNT for census family size), but it does not matter from which census family of the household they belong, because they will only be used if in fact there is just one family in the household; therefore select the first family in the household for simplicity. A counter must be included in the merge process in order to record the number of families in each household—specifically, the number of distinct values of the family identifier CF_ID for each value of HH_ID. After the output file is completed, the values of the new 'household type' variable can be derived in 'if-then' statements as follows:

- if the counter equals 2 or more, the value of the new household type variable is 'multi-family household';
- if the counter equals 1, then the value depends on the CFSTRUCTSIMPLE value of the one family retained. But it also depends on whether additional persons were found in that one-family household. For this, if the household size, NUNITS, is greater than the census family size, CFCNT, then additional persons are present, otherwise they are not; and
- if the counter equals 0 and NUNITS equals 1 then it is one person living alone; if the counter equals 0 and NUNITS is greater than 1 then it is a non-family household of two or more persons.

4.3 Selecting one person per family or per household

Some of the RDC and Hierarchical PUMF code books recommend which variable values to use as selection criteria for choosing one person-record to represent the household or the family. Any variable which has a value that always applies to one and only one person of each family or household can be used. Therefore, for the 2006 and 2011 RDC files or Dissemination database, one person per **household** can be selected using PERSNO=1 or HMAIN=1 (HMAIN=3 in 2006). In the Hierarchical PUMF, use PRIHM=1. For 1996 and 2001, use PERSNO=1 or HHPTR=0 in the RDC files. For the **family**, the variables CF_RP and EF_RP can be used to select one reference person per census family or economic family, respectively, back to 1996. Alternatively, identifier variables can be used if they are unique across all records in the file (see Section 5.3 for more information).

It is possible to apply the selection criteria (select units) so that there is a record for each economic family and each person not in an economic family, or for each census family and each person not in a census family. The variables to use are still CF_RP and EF_RP. As well, the variables for census family status or economic family status always have a category for persons not in a family, so they can be used to identify this population.

5. Other technical aspects of the census and NHS microdata

Users of the census and NHS microdata files will have to decide what population is applicable for their variables, what weight is appropriate to apply and whether the data need to be manipulated using identifier variables. This section provides additional details on these aspects.

5.1 Selecting the population for family analysis (the 'universe')

In the census, the applicable population or 'universe' for basic demographic characteristics—namely age, sex and marital status—is the 'total population'. This is the full target population of the census. However, since the 1976 Census, characteristics pertaining to families are not published for persons living in collectives.[13] When collectives are excluded, the resulting universe is the population in private households.[14]

Users of the National Household Survey RDC file do not need to actively restrict their study population to that of private households because the NHS had only private households as its target population. Also, the census public-use files of 1976 to 2006 exclude all collectives for simplicity.

Users of the Census Program's Dissemination database (either 100% data or sample data) or the RDC files from 1976 to 2006 (sample data) must restrict their analysis to private households or persons in private households if they are including any family-related or housing variables in their study. This is because, in the case of 100% data files, there are records for the total population—people in private households plus collectives—and in the case of the sample data files from 1976 to 2006, there are records for the non-institutional population—people in private households and non-institutional collectives.

Variable DOCTP serves to select the universe of private households or persons in private households.[15] For 2011, use DOCTP=1. In the NHS files, DOCTP=2 and 9 cover the private population, and there are no other records in these files. For 2006 back to 1991, use DOCTP=7 and 8. For 100% data for those years (Dissemination database only), also include DOCTP=6.

Analysts can refer to published census and NHS tables to validate their tables produced using Statistics Canada's Dissemination database or the RDC microdata files, provided that the results are for the same universe—that is, whether it is the total population, the non-institutional population, or the population in private households. Also, if the tabulated characteristics were collected on both the short questionnaire and long questionnaire, then it should be noted which of the two possible sources (100% data or sample data) was used to produce the particular table. The short-form characteristics are age, sex, marital status, family characteristics[16] and certain language characteristics. As a general rule, tables on these characteristics are published using 100% data unless certain long-form characteristics are also being shown in the same table (or the same set of tables). Full data and weighted sample data do not always give identical counts—even for the same universe, as explained in Section 5.2 on weighting. It is important to emphasize, however, that almost all analytical conclusions based on these data would be the same.

13. For housing characteristics, collective dwellings have been excluded since 1971. In 2011, people living in all collective types as their 'usual place of residence' made up about 2% of the total population.
14. Statistics Canada employees using the Dissemination database will notice that private households outside Canada, as defined in descriptions of the census target population, also account for a small portion of the total population—0.02% in 2011. The census includes them in tabulations of the total population or the non-institutional population (and they are associated with geographic locations in Canada via information that they provide); however, since 1976, it excludes them from the universe of 'private households'.
15. In the 100% Dissemination database files, variable IRIND is used to select the non-institutional population.
16. For the census years 1996, 2001 and 2006, only sample family data were available from the census. The published family tables were all produced on sample data for those years.

5.2 Applying weights to the sample microdata files

This section provides a summary of the final weight variables that analysts need to use in data tabulations of 1996 to 2011, whether for households, persons or family counts. For more information on weighting, refer to one of the PUMF user guides of 2011—either Chapter 3 in the Individuals PUMF or Chapter 4 in the Hierarchical PUMF.[17]

The census long-form databases and the 2011 NHS database are all sample files. For data extractions using these files, it is necessary to apply a weight variable to the records so that they represent the original target population from which they were selected. Accordingly, for 2006 and earlier years, the long-form census is weighted so that it sums to the counts of the short-form census (referred to as the 100% data), excluding institutional residents. The 2011 NHS is weighted so that it sums to the 2011 Census counts (100% data) for the universe of private households.

Although sample weighting can usually make the sample data sum to the control population at the total level for Canada, some minor differences will still exist between the count of any sub-population based on sample data and that from the same sub-population using the control file. This is true even after calibration of the weights based on common characteristic variables and geographic areas.[18]

In the Hierarchical PUMFs (2006 or 2011) or the Individuals PUMFs (any year), there is only one survey weight variable, called WEIGHT. It is applicable for tabulations at any level, whether persons, households or families. Select one person per household or one person per family and apply WEIGHT to those records in order to obtain weighted counts of households or families.

In the Dissemination database or the RDC files of 1996 to 2011, the weight variables are COMPW1 for the household and COMPW2 for the person. In fact these two variables almost always have the same value; that is, the value for the household, COMPW2, is equal to the value of COMPW1 for each person within the same household. This means that either variable can be used to weight households or persons. Either weight variable can also be used to weight families.[19]

Several 'replicate weight' variables have been provided in the public-use files since 2006. They serve a different purpose from the survey weight variables. Analysts can use them to estimate the sampling variability of their estimates, for example by calculating the coefficient of variation and producing confidence intervals. More information about the replicate weights and how to use them is provided in the PUMF user guides.

17. For the censuses before 1996, several additional points are necessary regarding the application of weights in order reproduce published counts for families and households. The explanations for 1991 can be found in Roberts (2012).
18. Illustrations and explanations of these differences can be found on-line in the *Sampling and Weighting Technical Report* (Statistics Canada, 2015). Similar reports exist for earlier censuses as well.
19. There is just one case where COMPW1 does not equal COMPW2 and it does not affect the universe of private households; therefore it does not affect analysis of families or housing characteristics. In brief, COMPW1 was set to 'missing' for certain collectives in 2006, in the RDC file only, as there was an issue with the classification of some seniors' residences. As a consequence, it is necessary to use the person weight, COMPW2, to weight the non-institutional population in 2006.

5.3 Manipulating multilevel data using the identifier variables

It was explained in Section 4.4 that one person per household or family can be selected from a hierarchical person-level file using appropriate selection criteria. Identifiers that are unique for all units in the file, which is the case for HH_ID, PP_ID, CF_ID and EF_ID, can be used for this purpose. The user can sort the person file by the household identifier (or a family identifier) and then retain only the first person associated with each new value of that identifier.

The following examples of programming illustrate how identifiers can be used to create a family- or household-level file from the hierarchical file (Example 5.3.1), to create new multi-level variables (Example 5.3.2), or to even create a single hierarchical file from a relational database of multiple files (Example 5.3.3). SAS code is shown because it is commonly used at Statistics Canada. However, even if the programming syntax is not familiar (or complete), the explanations should give some idea of the use of identifiers for multi-level data manipulation. Note that any fictitious file names and variable names are shown in italics. Options such as the variables to keep or drop from the original file are not indicated, for simplicity.

Example 5.3.1: Create a 'census family' file from the hierarchical (person-level) file.

The file has to be initially sorted in order of the census family identifier, CF_ID.

```
PROC SORT DATA=PPFILE; BY CF_ID; RUN;
```

Next, one person per census family can be selected with an 'IF' statement within a data manipulation statement.

```
DATA NEWFILE; SET PPFILE;
  BY CF_ID;
  IF FIRST.CF_ID;
RUN;
```

As the program looks through all the person records in order of their values of CF_ID, it will retain each record that has a 'new' value of CF_ID compared to all the previous records it has reviewed. Note that the records of non-family persons will be retained in the new file along with one person per family, because each non-family person has a different value of CF_ID (that is also different from all family persons). To exclude non-family persons if desired, add a second condition that uses any variable that allows them to be identified, such as C_FAM=0 or CFAMST=(the value for a person not in a census family).

```
DATA NEWFILE; SET PPFILE;
  BY CF_ID;
  IF FIRST.CF_ID AND CFAMST NE (value for person not in a census family);
RUN;
```

where 'NE' signifies 'not equal to'.

To re-use this example for economic families, replace CF_ID and CFAMST with EF_ID and EFAMST.

To create a file of all households, select one person per household: IF FIRST.HH_ID.

Example 5.3.2: While creating an 'economic family' file, create a counter of the number of foster children per family.

First exclude the persons who are not in economic families. All foster children are members of an economic family.

```
DATA NEWFILE1; SET PPFILE;
  IF EFAMST NE (value for person not in an economic family);
RUN;
```

Next, the file has to be sorted by the family identifier, in this case EF_ID (not shown). The last person with a given value for the identifier (family or household) has to be used when deriving either a counter variable such as the number of foster children or an indicator variable such as yes/no or true/false for any characteristic. In the SAS example below, the name of the new variable is NUMFOSTER. It has to be initialized to 0 at the start of each new family as indicated by a new value of EF_ID (IF FIRST.EF_ID).

```
DATA NEWFILE2; SET NEWFILE1;
  BY EF_ID;
  RETAIN NUMFOSTER;
  IF FIRST.EF_ID THEN NUMFOSTER=0;
  IF EFAMST=(value for foster child) THEN NUMFOSTER=NUMFOSTER+1;
  IF LAST.EF_ID THEN OUTPUT;
RUN;
```

In the example above, to count foster children by households instead of economic families, use HH_ID in place of EF_ID. EFAMST can be re-used as the indicator of foster children at the person level.

Example 5.3.3: Re-create the person-level hierarchical file from separate family, household and person-level files.

This example is included because it illustrates the use of identifiers to merge files of different units of analysis. It may also help to illustrate the structure of the RDC file and the hierarchical PUMF.

Three data manipulation steps are used, which keeps it simple because only two files are merged at a time. In each step, the two input files have to be sorted by the identifier that is common to them. In order to keep only persons at the end, the program asks (in an IF statement) to retain only the records found in file 'A'—the value which is given to the person-level file.

Step 1: Start with the person file and create a new file that attaches the household variables to each person record, sorting first.

```
PROC SORT PPFILE; BY HH_ID; RUN;
PROC SORT HHFILE; BY HH_ID; RUN;

DATA NEWFILE_PPHH; MERGE PPFILE (IN=A) HHFILE;
  BY HH_ID;
  IF A;
RUN;
```

Step 2: Add the economic family variables, sorting first.

```
PROC SORT NEWFILE_PPHH; BY EF_ID; RUN;
PROC SORT EFFILE; BY EF_ID; RUN;

DATA NEWFILE_PPHHEF; MERGE NEWFILE_PPHH (IN=A) EFFILE;
  BY EF_ID;
  IF A;
RUN;
```

Step 3: Add the census family variables, sorting first.

```
PROC SORT NEWFILE_PPHHEF; BY CF_ID; RUN;
PROC SORT CFFILE; BY CF_ID; RUN;

DATA NEWFILE_PPHHEFCF; MERGE NEWFILE_PPHHEF (IN=A) CFFILE;
  BY CF_ID;
  IF A;
RUN;
```

Note that this is a partial example only. One other step to consider is how to avoid null values for non-family persons, i.e., persons who do not have any corresponding record in the family file. Such records may have a value of CF_ID, but the same value of CF_ID cannot be found in the census family file, so no link is made. They will have the census family variables attached to them because they are loaded for all records simultaneously, but there will not be values for persons not in a census family. To avoid this, the family variables could be initialized in the new file using a standard 'not applicable' value.

Summary

The availability of census and NHS microdata files in the form of the Census Program's Dissemination database, in the RDCs, as well as public-use files, provides researchers with the opportunity to do in-depth analysis on a variety of topics, often involving smaller populations and lower levels of geography. This article has provided the basic information needed by researchers in order to understand and use family-related variables in their analysis. Several specific examples illustrated how variables can be created from different levels of data, that is, person, family or household. Also included was a discussion of the demographic and family variables, technical features like the structure and organization of the data files, and the use of weights and identifier variables.

This article is a useful supplement to technical documentation already provided with the 2011 microdata files and those of previous censuses. It may be of interest to researchers who want to extend their study of other topics covered by the census, for example, by looking at labour market situations and incomes at the family level and not just the person level. The information that applies to the census may also apply in varying degrees to other surveys of Statistics Canada that cover census family and economic family data.

References

Bohnert, Nora, Anne Milan and Heather Lathe. 2014. "Enduring diversity: Living arrangements of children in Canada over 100 years of the census", *Demographic Documents*, no. 11, Statistics Canada Catalogue no. 91F0015M.

Peller, Peter. 2012. "The 2006 Canadian Census Hierarchical PUMF", *DLI Update*, Statistics Canada, volume 13, no. 1.

Roberts, Georgia. 2012. "Analyzing Census microdata in an RDC: What weight to use?", *The Research Data Centres Information and Technical Bulletin*, Statistics Canada Catalogue no. 12-002-X, volume 5, no. 1.

Statistics Canada. 2015. *Sampling and Weighting Technical Report, 2011 National Household Survey (NHS)*, Catalogue no. 99-002-X2011001.

Statistics Canada. 2012. *Census Dictionary, 2011 Census*, Catalogue no. 98-301-XIE.

Statistics Canada. 2012. *Families Reference Guide, 2011 Census*, Catalogue no. 98-312-XWE2011005.

United Nations. 2015. *Principles and Recommendations for Population and Housing Censuses, Revision 3*, New York.

Appendix: List of demographic and family variables

Table A1 shows the variables for 2011 that contain some demographic or family-related content. The code books accompanying the RDC and public-use microdata files include the full variable descriptions and variable categories. For the Dissemination database, this information is found in E-dict.

Variables beginning with 'CF' apply to the census family concept, and variables beginning with 'EF' apply to the economic family concept.

The variables indicated with an asterisk '*' are also available in the Individuals PUMF of the 2011 NHS, although the names may be different: for example, CFCNT_PP is called CFSIZE in the PUMF and EFCNT_PP is called EFSIZE. The categories may not be identical in the PUMF because they have been grouped to show less detail for confidentiality purposes, or the information from two variables has been combined for the same purpose (for example, variables CFKID1 and CFKIDLT1 are provided as a single variable PKID0_1 in the PUMF). .

The Hierarchical PUMF of 2011 contains the same demographic variables as the Individuals PUMF (age groups, sex and marital status) but fewer distinct age groups. It has a slightly different selection of family-related variables: CF_RP, EF_RP, CFSTATSIMPLE and CFSTRUCTSIMPLE. However, due to the presence of identifiers (HH_ID, PP_ID, CF_ID and EF_ID), it is possible to derive much more information on families and households from the Hierarchical PUMF than from the Individuals PUMF, as was explained in Section 2.

Table A1
Demographic and family variables, 2011 Census and 2011 NHS

Level	Short description	Variable names	
Person level	Demographic variables on age, sex and date of birth	AGE	DAYB
		AGECONT	MTHBD
		AGEGR5 *	BRTHYR
		SEX *	
	Marital status (legal and de facto)	MARST	COMLAW
		MARSTH *	HWCLPR
	Relationship to Person 1 of household	R2P1	
	Reference person (allows selection of one person per family)	CF_RP	EF_RP
	Status of person in family and in household	CFAMST CFAMSTDET CFAMSTSIMPLE	CFSTAT CFSTATSIMPLE *
		EFAMST EFAMSTSIMPLE EFAMST06	
	Family size (1 for persons not in a family)	CFCNT_PP or CFSIZE*	EFCNT_PP or EFSIZE*
	Number of children of person, by ages, and youngest child	PRESCH0T18	PRESCHNUM
		PRESCHAGEMINGR	PRESCHSET7A
		PRESCHILD	
Family level: census family (CF) or economic family (EF)	Family structure (or type)	CFSTRUCT CFSTRUCTSIMPLE	EFSTRUCT EFSTRUCTSIMP
	Number of children in family, by ages, and youngest child	CFKID0T14	CFKID2T5 *
		CFKID0T18	CFKID5T9
		CFKID0T4	CFKID6T14 *
		CFKID0T5	CFKIDAGEMINGR
		CFKID1 *	CFKIDGE25 *
		CFKID10T14	CFKIDLT1 *
		CFKID15T17	CFKIDNUM
		CFKID15T24 *	CFKIDS
		CFKID18T24	CFKIDSET7A
		CFKID2	
	Characteristics of first or second spouse/partner or parent in family	CFAGE1STPRSN	CFSEX2NDPRSN
		CFAGE2NDPRSN	EFAGE1STPRSN
		CFMAR1STPRSN	EFAGE2NDPRSN
		CFMAR2NDPRSN	EFSEX1STPRSN
		CFSEX1STPRSN	EFSEX2NDPRSN
Household level	Number of generations and type of middle generation	FAMGENSTAT	FAMMIDSTAT
	Number of persons in, or not in, a family, in the household	CFHH	CFNM

*refers to variables available in the Individuals PUMF of the NHS. This file also has variables for household size (HHSIZE, called NUNITS in the RDC file) and household type (HHTYPE); however, it does not include CFSTRUCTSIMPLE.