The 2006 Canadian Census Hierarchical PUMF

Peter Peller, University of Calgary

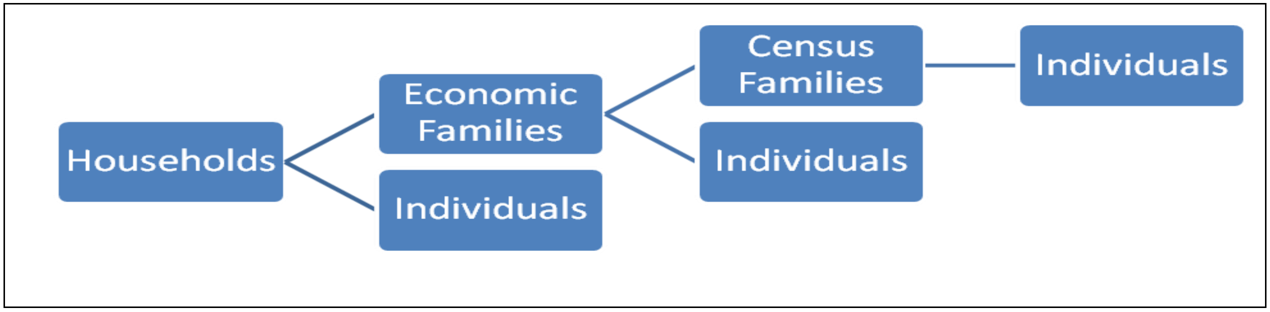
This year for the first time, Statistics Canada released a census hierarchical public use microdata file (PUMF). Although PUMFs have been released for each census since 1971 (Boudreau & Manriquez, 2006), the 2006 hierarchical PUMF is different from the separate individuals, families and households PUMFs previously disseminated. It is important for data librarians to understand this unique file so that they can be more knowledgeable about when and how to use the file. This article will examine the concept of hierarchy as it pertains to microdata files and will discuss the specific features of the 2006 Census Hierarchical PUMF.

In the context of PUMFs, hierarchy refers to a relationship between affiliated data records at different levels or units of analysis. This relationship is structured as a descending rank network with a one-to-many association from the higher to lower level records. Figure 1 illustrates the possible relationships in the 2006 Census Hierarchical PUMF where the descending rank is private households in private occupied dwellings, economic families, census families and individuals. The *2006 Census Dictionary* (2008) has the following definitions for these levels:

> **Household, Private** – Person or group of persons occupying the same dwelling
>
> **Economic Family** – A group of two or more persons who live in the same dwelling and are related to each other by blood, marriage, common-law, or adoption
>
> **Census Family** – Refers to a married couple (with or without children of either or both spouses), a couple living common-law (with or without children of either or both partners) or a lone parent of any marital status, with at least one child living in the same dwelling

In some cases the hierarchy is simply households containing one or more unrelated individuals; this occurs where the individuals do not comprise an economic family. By definition, some economic families do not contain census families, but all census families are contained within economic families.

Figure 1. Census Household Hierarchy



Hierarchical PUMFs fall into one of two categories: separate and combined (Watkins, 2001). Separate hierarchical files are a group of related files where all the records in a specific data file are at the same unit of analysis, but at a different unit of analysis from the records in the other data file(s) (see Figure 2). A combined hierarchical file is a single data file containing related records at different units of analysis, or in other words, a data file with different levels of information on different records (see Figure 3).

These files contain “key variables” (Dupriez & Boyko, 2008) that provide links between the related records in different files (in the case of separate hierarchical files) or between the related records within a single file (in the case of a combined hierarchical file). Key variables are also referred to as “unique identifiers” and are usually numeric. In Figures 2 and 3 the Household Identifer (HHID) is the key variable that links the individuals to their respective households.

Figure 2. Separate Hierarchical Files

Household File

Household ID	*Region*	*Tenure*
1	Quebec	Own
2	BC	Rent

Person File

Household ID	*Person ID*	*Gender*	*Age*	*Marital Status*
1	11	Male	40-44 years	Common-Law
1	12	Female	40-44 years	Common-Law
2	21	Male	35-39 years	Married
2	22	Female	30-34 years	Married
2	23	Male	0-4 years	Single
2	24	Female	5-9 years	Single

Figure 3. Combined Hierarchical File

HHID = 1	Quebec	Own	
PID = 11	Male	40-44 years	Common-Law
PID = 12	Female	40-44 years	Common-Law
HHID = 2	BC	Rent	
PID =21	Male	35-39 years	Married
PID = 22	Female	30-34 years	Married
PID = 23	Male	0-4 years	Single
PID = 24	Female	5-9 years	Single

Statistics Canada has produced hierarchical PUMFs before. Some examples are the Canadian Travel Survey PUMF with related person and trip files, and the General Social Survey (GSS) PUMFs with related main and incident (or episode) files. In the case of the Canadian Travel Survey (CTS), there are two separate hierarchical files: the main file contains variables describing all the respondents, and the trips file contains variables describing all the trips that the respondents have taken. In the CTS the “person identifier” variable serves as a key variable that links the trips records to the corresponding persons.

Until the release of the 2006 Census hierarchical file, census PUMFs only had, what could be termed, “partial hierarchy” (Boudreau & Manriquez, 2006). They were basically flat files that contained records at a specific unit of analysis; most of the variables describing those records were at the same level but a few of the variables described other affiliated units of analysis. For example, on the individuals file each record had some matching family and household variables such as family income, number of persons in household, etc.; similarly, the families and households files each had some matching individual variables such as total income of female spouse, ethnic origin of the primary household maintainer, etc. What made it a partial hierarchical file wasn't so much the scarcity of variables at the other units of analysis, but rather the fact that not all connected records were present and/or identifiable. In the case of the individuals file, all the individuals from one household were not necessarily present in the sample and there was no unique identifier that would enable a user to link the individuals from the same household.

According to the *Documentation and User Guide*, the 2006 Census Hierarchical PUMF contains non-aggregated data for an "approximately 1% sample of the population in private households in private dwellings in Canada" and provides information on the "demographic, socio-cultural, ethno-cultural, housing and economic characteristics of census families, economic families and private households as well as the individuals that define these units" (Statistics Canada, 2011). The key point here is that this file contains all the individuals who make up the households in this sample and each individual is linked to their corresponding family and/or household. This is what makes the file truly hierarchical.

The data structure of the 2006 Census Hierarchical PUMF is "rectangular." Combined hierarchical PUMFs have different levels (units of analysis) of data "nested" as separate records as in Figure 3; however, single hierarchical files can also be "rectangularized" meaning that the data from all levels of the hierarchy have been combined at one level (see Figure 4) in essence creating a flat file (Rafferty & Wathan, 2008). This simplifies the data structure, but the trade-off is that household and family level data are repeated for each matching individual which requires more space.

Figure 3. Rectangularizied Hierarchical File

Household ID	*Person ID*	*Region*	*Tenure*	*Gender*	*Age*	*Marital Status*
1	11	Quebec	Own	Male	40-44 years	Common-Law
1	12	Quebec	Own	Female	40-44 years	Common-Law
2	21	BC	Rent	Male	35-39 years	Married
2	22	BC	Rent	Female	30-34 years	Married
2	23	BC	Rent	Male	0-4 years	Single
2	24	BC	Rent	Female	5-9 years	Single

As with other data samples, the use of the weighting factor is necessary to obtain population estimates. The weighting of the census hierarchical file has been designed in such a fashion that the individuals weighting factor actually applies to all units of analysis. The *Documentation and User Guide* has detailed procedures on how to estimate frequencies, ratios, proportions, and averages using this file. It also covers in depth, the estimation of the sampling variability or coefficient of variation from which researchers can calculate a confidence interval for their estimates (Statistics Canada, 2011).

Hierarchical data is much more complex to anonymize than data at one level (Boudreau & Manriquez, 2006). The reason for that is "individuals by themselves, may not stand out, but through association with some other individuals in the household, they may make the household unique even when the geographic detail is greatly reduced" (Hawala, 2003, p. 7). To protect the confidentiality of the respondents in this file, Statistics Canada (2011) has taken a number of actions such as aggregating industry and occupation categories, combining the rented and band housing into one category, random rounding and top-coding of income and shelter costs, limiting households to "seven or more persons", and reducing geographic detail to regions (British Columbia, Prairies, Ontario, Quebec, Eastern Canada, Northern Canada). The 2006 Hierarchical PUMF has 124,358 households as compared to the 2001 Households PUMF which had 312,513 households. This smaller sample of households in the 2006 Hierarchical PUMF also reduces the chance of identifying someone, because each entity represents more of the universe when weighted. All the disclosure protection actions are described fully in the *Documentation and User Guide* (Statistics Canada, 2011).

There are several benefits associated with the creation of the census hierarchical PUMF. As is the case with other microdata files, this file will allow researchers to group and manipulate variables to create tabulations that have been excluded from other census aggregate products (Statistics Canada, 2011). Also, the census hierarchical PUMF brings Canada into line with the output from other international censuses and enables multinational analyses (Boudreau & Manriquez, 2006). Although the 2006 Census Individuals PUMF is usually more appropriate to use (larger sample and more variables) when looking at individuals only, the major benefit of a hierarchical file, however, is that it allows for examination of cross level effects at both the individual and group levels at the same time (Rafferty & Wathan, 2008). Researchers will find this ability to use information from one level to augment the analysis at another level very valuable.

Although it was a long time coming – due to the extensive efforts to protect the privacy of the respondents – the 2006 Census Hierarchical PUMF was definitely worth the wait. With its hierarchical structure it opens up new possibilities for researchers. We, as data librarians, can contribute to its success by being a well-informed link between this PUMF and the researcher.

References

Boudreau, J., & Manriquez, R. (2006). *Research into the possibility of releasing hierarchical public use microdata files for the census of population*. Statistics Canada. Retrieved from http://www.fcsm.gov/03papers/Manriquez-Boudreau.pdf

Dupriez, O., & Boyko, E. (2010). *Dissemination of microdata files: Principles, procedures and practices.* (IHSN Working Paper No 005) International Household Survey Network. Retrieved from http://www.ihsn.org/home/index.php?q=activities/working-papers

Hawala, S. (2003). Microdata Disclosure Protection: Research and Experiences at the US Census Bureau. Retrieved from http://www.census.gov/srd/sdc/microdataprotection.pdf

Rafferty, A., & Wathan, J. (2008). *Working with survey files: Using hierarchical data, matching files, and pooling data.* Economic and Social Data Service. Retrieved from http://www.esds.ac.uk/government/docs/workingwithsurveyfiles.pdf

Statistics Canada. (2011). *2006 census public use microdata file (PUMF), hierarchical file: Documentation and user guide.* (Catalogue no. 95M0029XVB). Retrieved from http://equinox.uwo.ca/docfiles/2006_Census/pumf/hier/pumf%20user%20guide.pdf

Statistics Canada. (2008). *2006 Census dictionary*. (Catalogue no. 92-566-XWE). Retrieved from http://www12.statcan.ca/census-recensement/2006/ref/dict/index-eng.cfm

Watkins, W. (2001). *Complex files: Pasting and cutting with SPSS.* (CAPDU/DLI Ontario and Quebec Training). Retrieved from https://ospace.scholarsportal.info/handle/1873/216