Notes on Census/NHS 2011 for Updating SuperDemographics 2014

Zhen Mei, Ph.D.
Manifold Data Mining Inc.,
December 09, 2014

Census 2011 was conducted on May 10th, 2011 with a mandatory short form of Census Questionnaire and a voluntary National Household Survey (NHS), which replaced the historical mandatory long form of census questionnaire.

The mandatory short form consisted of 10 questions and was required by law to be filled by all Canadian households. The national response rate of the short form was 98.1%, the highest in the Census history. Ontario and Prince Edward Island had the highest response rate at 98.3%, while Nunavut had the lowest response rate at 92.7%. Data collected from the mandatory form of Census 2011, i.e., population age group, family structure, household composition, dwelling types and languages were released completely by Statistics Canada ("StatCan") at the DA (Dissemination Area) level. We were able to use these data to fully update the population, family, dwelling, household, home language, mother tongue categories in the SuperDemographics 2014. To capture the undercount of population in the Census, estimate current year and project future population statistics, we considered the StatCan's post-Census survey, analyzed and incorporated the historical trends, regional birth and mortality rates, migration and immigration statistics, real estate statistics and new postal code information as well as directory books. We used a bottom up technique to aggregate key data from the ground/street level to the standard Census geographies. At the same time, we also took a top down approach by comparing StatCan's research and publications on population estimates and projections with our results in order to identify abnormalities and establish consistency. In this way we leveraged the strength of data both at street and municipal levels.

Instead of the mandatory long form questionnaire on ethnicity, immigration, labour force, income, dwelling value, religion…, government decided to change it into voluntary in Census 2011. Thus StatCan conducted the voluntary National Housing Survey ("NHS") within four weeks of the May 2011 Census among 4.5 million Canadian households. Unfortunately the response rate was merely 68.6 percent, much lower than the traditional 93.5 percent when the long form was mandatory. Prince Edward Island had just 60.3% responders. Saskatchewan's response rate was 63.8% although it has been one of the fastest growing regions in Canada for the last 10 years. Even worse, the variation in response rates by geography and by questionnaire is big and irregular. Unlike the previous Census in which response to the mandatory long form was also biased, but more in a systematic fashion and can be corrected statistically, the current voluntary NHS survey created a "random" variation which is much hard to fix for small geographies. StatCan put extraordinary efforts and resources into reconciliation of gaps and aberrations in the NHS data, although nearly half of Statistics Canada's employees were notified in 2012 that their jobs might be eliminated as part of austerity measures

imposed by the federal government. StatCan suppressed data for 25% Census Sub-Divisions ("CSD") in their final publication of NHS data due to poor quality. Data for thousands of smaller communities were excluded from the release because it was not considered reliable. For Saskatchewan, only 57% CSDs were covered in the publication of data on aboriginal peoples, immigration and ethnocultural diversity. Numbers of data attributes on of education, labour force, employment, migration, income and housing were reduced drastically. Moreover, StatCan decided not to publish DA level data because of high and irregular non-response rate. Given all these deficiencies, the NHS is still the most reliable and systematic survey about Canadian culture, labour force, income and housing.

In addition to the CSD level NHS data, we purchased a custom tabulation of DA level NHS data from StatCan. We analyzed patterns of non-response rates, data-mined systematically the cohesive structure (e.g., relationship between ethnic origins and religions) of the NHS data, established links to the Census (short form) data and incoporated the historical trends. We recognized seven major categories of deficiencies in NHS data at the DA level:

1. No data at all for DAs where the non-response rate was too high or the number of population was too low.
2. Some variables are completely suppressed or dropped, while other variables are available and complete for some DAs.
3. Some variables are only partially available. Part of values of the variables are suppressed.
4. Round-off of variables by different integers 5, 10, or 50, e.g., population by 5, ethnicity by 10 and income partially by 50.
5. Inconsistence between responses of the short form questionnaires and NHS due to the nature of two separate surveys.
6. Inconsistence between DA level and CSD level data because StatCan applied imputation of missing values to the CSD level data based on the historical Census 2001 and 2006 which may have in-compatible coherent structure for some CSDs., while DA level data is the direct aggregation of responses.
7. Inconsistence among variables within same category. Due to the voluntary nature of the NHS survey, responders may have filled in a main question but not questions in the sub-category, or vice versa, e.g., university education and majors of study.

To improve the DA level data, we considered the historical trends and applied the nearest neighbourhood techniques to fill in the missing values for DA without data. For DA's with partial data we used the spatial regression technique to fill the unknown value of the variables. We overlaid the NHS data with taxfiler's data, labour force survey and third party survey like the Return-To-Sample survey from the Bureau of Broadcast and Measure which we have accumulated over one million responders stratified across Canada. We filled the gaps in individual variables where StatCan suppressed the values due to low response rate. We adjusted the Census and NHS data and eliminated the

round-off errors. We also validated and confirmed the consistence between our DA level estimates with StatCan's CSD and FSA (Forward Sortation Area, i.e., first three digits of a postal code) level data. We applied the following three approaches to refine our models of estimates and projections:

1. Hold-off geographic areas to verify and refine variation in different regions across Canada.
2. Use historical Census to build models and current Census to verify and adjust the methods.
3. Compare with housing permits, real estate statistics and CPI (Consumer Price Index) to capture dynamics of the population and economic statistics.

We completed our updates of SuperDemographics with high accuracy and consistence with historical trends, current StatCan's estimates at municipal and provincial levels and 2014 October Labour Force Survey statistics.

We look forward to your comments and welcome your feedbacks helping us improve our data quality. Please contact us if you have any questions.

Zhen Mei, Ph.D.
Principal, Analytics and Modelling:
Zhen@manifolddatamining.com
Manifold Data Mining Inc.
220 Duncan Mill Road, Suite 519
Toronto, ON M3B 3J5
Tel: 416-760-8828
Fax: 416-760-8826