Data Purchase and Access Working Group meeting
September 28, 2016
1:30 – 2:30pm Eastern time
https://cdp.adobeconnect.com/theboardroom
Teleconference number: 1-888-271-3643

PRESENT
Isabelle Lépine - Montreal
Auburn Larose - Wellington-Dufferin-Guelph
Chelsea Turan – Simcoe County
Andrea Dort - Peel
Valentyn Kliuchnyk - York
Natalie Hui - York
Jasmine Ing- Calgary
Cheryl Hitchen - Kingston
Louisa Wong – Hamilton
Heath Priston – Toronto
Ted Hildebrandt – Halton

AGENDA

1. Methodological issue with Income Inequality data
2. Finalized Schedule B
3. Equifax data
4. ePCCF and our licensing issues
5. Custom geographies
6. General Social Survey
7. Labour Force Survey
8. Taxfiler custom tabulations
9. Other business
10. Next meeting

1. Methodological issue with Income Inequality data

It was noted that the CDs of Halton (3524) and Toronto (3520) had their aggregate income suppressed for both the first and the tenth deciles in the 2013 data table. The reason that was provided for this suppression was a dominance rule:

Dominance for top and bottom deciles
"Our suppression rules for dominance dictates that if one value within a data cell (extremely low or extremely high) has too much of an impact as compared to the other values within that cell, then that cell is suppressed. A good example of this would be if we have one extremely rich individual in a small rural region. For example, if you have 19 individuals the top decile and the individuals in that deciles have incomes around $40,000, but the 20th individual has an income of a few million dollars, then the cell gets suppressed. This is done in order to protect the confidentiality. Someone familiar with the area who knows that only one individual there truly has a high income would be able to breach the confidentially that we guarantee to our survey respondents or, in the case T1FF, tax filers.

...We also have a secondary rule which triggers supplementary suppression in order to help protect confidentiality. When one decile is suppressed, a second decile is automatically suppressed. Hence if the 10th decile is suppressed for dominance associated to an income amount, and no other decile is initially suppressed, the first deciles will automatically be suppressed."

It was determined that given the Top 5% and Top 1% were not suppressed, that the dominance rule was being applied to the 1st decile and that the 10th decile was subject to the supplementary suppression.

One way around this problem would be to bottom-code the first decile so that negative incomes are coded as $0. These negative incomes would not dominate the first decile and then neither decile would need to be suppressed.

The question put to the group was: do we re-order the tables using this bottom-coding and proceed with future orders using this methodology?

Valentyn (York): York has purchased these data previously and there are two options: exclude the negative income earners completely or count their incomes as zero. See York's 'Income Inequality Trends in York Region - 1997 to 2012'.
- The group decided that it was best not to exclude the negative income earners so as not to disrupt the decile structure.

Jasmine (Calgary): It would be of interest to Calgary and likely to other communities that fall on tough times to not have the negative incomes of the first decile suppressed, to be able to accurately assess the impact of economic downturns.

ACTION ITEM
As a first course of action the CDP team will get a cost estimate to have the 1st decile bottom-coded and consider having the table re-run to produce two copies of the 1st decile, one using the current methodology and the second bottom-coded to count negative incomes as $0. The CDP team will look at the cost of carrying out this re-production for just the 2013 data and also for previous years.
The CDP team will look at other geographies to see what the impact is on the aggregate data – i.e., how does this methodology affect the data?

A second issue came to light while the working group was examining the Income Inequality data which was that it was noted that in certain geographies, 'higher-level' deciles had lower aggregate incomes than 'lower-level' deciles.

ACTION ITEM
The CDP team will investigate this issue to determine if it is an error or an irregularity caused by the methodology.

UPDATE: Regarding this error, Statistics Canada has provided the following response with an example data table that will be included in an email with this document:

"Let's assume we have N = 200 observations in a group (a specific CSD) with income values ranging from Value1….. to Value200. To create the deciles we divide N by 10 = 200/10 = 20 observations in each decile (this is in theory). However, in real life, the values from value1 to value200 are not all distinct. Some of them are repeated. When a repetition happens around 20, 40, 60, ….., 160, and 180 we

have to decide what to do. Some purists would base their deciles only on the count splits (20 in each decile), and in these case every "higher decile aggregate" would be greater are equal to any "lower decile aggregate". Some others will create their deciles based on the THRESHOLDS values at the cut-off (All the observations with the same value have to be in the same decile). In this later case the counts of observations in each decile will shift and won't be exactly the same. For this request, we adopted the second approach, and we are putting the observations with the same values equal to the threshold in the lower decile. This is why the counts in each decile are not always the same. Even for the counts that look the same (in the rounded data), they could have up to 9 observations differences (i.e.: 15 and 24 round both to 20). When we aggregate the values of these two final groups (mainly with low counts in each decile), this could create the situation of the "Aggregate" of a lower decile being higher than the "Aggregate" of a higher decile (i.e.: The Aggregate of the 24 observations of the lower decile being higher than the Aggregate of the 15 observations of the higher decile). Attached is an example showing how the phenomenon observed could happen (The aggregation of Decile 8 in the example is higher than the aggregation of Decile 9. The real income data has a level of repeated values a lot more frequent than the example given here).

If the client wants us to apply the first approach described above instead of the second, we can do it. But, in this case, the counts by decile would be meaningless (they will be exactly the same). All what would be needed is a "Total count" or a "Count by decile".

ACTION: The CDP team will discuss with the DPAWG as to the preferred methodology.

2. Finalized Schedule B

The finalized Schedule B has been posted. Questions regarding the document can be directed to the CDP team.

3. Equifax data

We have received, processed and posted the Equifax data to the catalogue. We are very much looking forward to hearing how these data are being used. The non-mortgage debt data contains many more variables than have been posted, including cuts by various types of debt products.

4. ePCCF and our licensing issues

We are not allowed to distribute the PCCF file in its entirety. Users will have to make requests for cuts of the file that are small enough in size to respect the license agreement. This same licensing issue exists with the Statistics Canada version of the file as the restrictions originate with Canada Post.

5. Custom geographies

The geographer at Statistics Canada is working on a cost estimate for geo-coding the custom geographies that have been submitted to date for the purpose of ordering taxfiler data using these geographies. There are still several consortia that are waiting to submit their geographies. It should be noted that Statistics Canada will be releasing geography files associated with the 2016 Census in November and that some consortia are waiting on these files before producing their custom geographies. Leads will be notified when we are ready to place an order.

Jasmine (Calgary): Would it be possible to post the shapefiles that were used to create the custom geographies as a reference for users?

The CDP would support implementing this idea. It would require some information or a methodology document from the 'owners' of the custom geographies in order to interpret them. Some organizations do not want to have their custom geographies accessible to the public.

6. General Social Survey

After some investigation into the quality of the General Social Survey data that we received it was decided that General Social Survey would be considered a low priority product this program year, as there are too few geographies of interest to the CDP without suppression or data quality warnings.

7. Labour Force Survey

We have obtained Labour Force Survey tables at the CMA and Economic Region level. We will evaluate this product's popularity when it comes time to purchase these data again next year. The data contain additional cross-tabs including wage rates by NOC and NAICS classifications and duration of unemployment.

8. Taxfiler custom tabulations

We have received some comments regarding the custom tabulations for taxfiler data, mostly regarding the data suppression that is likely for this order with small geographies:

Family Table 1 – add the following:
o Median income for Couple families with children (0-17)
No problem with this one.

Family Table 18 – add the following:
o Children & Youth: 0-6, 0-12, 0-17, 18-24, 25-29; Seniors: 55-65, 65+, 65-74, 75+ Yes, no problem except that these splits are going to generate a high volume of suppressions in the smaller areas.
o Add Families with children aged 0-17 (i.e., Couple Families with 0, 1, 2, 3+ children aged 0-17; Lone parent families with 0, 1, 2, 3+ children aged 0-17, etc.) Yes.
o Separate male and female lone-parent families Yes, this is doable, but could generate a lot of suppressions (mainly for male-Lone-parents) because of low counts in smaller areas.
o Separate male and female population aged 65+ Yes, this is doable, but could generate lot of suppressions because of low counts in smaller areas.

Family Table 6
o Source of income: Separate the Guaranteed Income Supplement from Old Age Security payments. Is it possible to break this down further?
Yes, this could be done. But if for any reason the combination of these two variables has been suppressed on the standard tables (That are already public); or any of the two parts is suppressed (for

any reason); both parts are going to be suppressed as a result. The suppressions will be done while keeping in mind that the standard tables data is already public.

NID Table 5 (A, B, C)
- Add a $1,000,000+ income category

Yes this is doable with the warning that this new category would be suppressed for all the small geographies and most of the medium size geographies.

Identify a wealth indicator

If the client has in mind some sort of Wealth indicator based on Tax data, we could discuss it and if doable we can implement it for him.

Since the meeting, Jasmine Ing suggested the Calculated financial assets data tables, for which we have requested a cost estimate for 2014 data at all available geographies.

9. Other business

Jasmine (Calgary) suggested sharing shape files of custom geographies (see above).

10. Next meeting

The next meeting is set for November 16th, 2016 at 1:30pm Eastern time.