Project for the Regional Advancement of Statistics in the Caribbean

On Adjusting for Dwelling Non-response in a Census

David Dolson, PRASC

March 16, 2020

Updated October 31 2022

This note is primarily directed towards a consideration of the treatment for dwelling level non-response in a de jure census of population. However, to fully place it in context some discussion is also provided on relevant aspects of data collection.

During data collection one of the enumerator's first jobs is to ensure her listing of dwellings is accurate. Depending on the particular census strategy, this may involve adding dwellings not appearing on the initial list, cancelling (or deleting) dwellings on the list which no longer exist (e.g. demolition) or which are not in fact dwellings (e.g. a business). Each dwelling requires a classification of its occupancy status on Census day.

This classification is difficult and errors are made by enumerators – some occupied dwellings are classified as not occupied and some not occupied dwellings are classified as occupied. In Canada, amongst the roughly two percent of non-response this classification error rate is about 12 to 20% in each direction! Unfortunately correcting for these errors requires estimation of the error rates by means of a sample survey in which expert recoding is done followed by modelling procedures applied to non-sampled dwellings. Instead we will assume the dwelling classification is correct.

Some censuses further identify some dwellings as being seasonally vacant. Whether persons are present or not at the time of enumeration or on census day, they never contain *usual* residents.

Another set of dwellings with no *usual* residents, are those with diplomatic or military staff of other countries.

Also, some censuses classify some non-responding occupied dwellings as *closed*. Such dwellings are those which the enumerator is able to confirm have usual occupants but they are absent throughout the data collection period (i.e. a long term absence). These dwellings should be treated as occupied non-responses.

Ultimately each occupied dwelling on the final list is either a respondent or a non-respondent and a decision must taken regarding what should best be done to account for the nonrespondents.

The Structure of Non-response

For the countries where PRASC has reviewed the information, all have similar means of uniquely identifying a dwelling – geo identifiers down to an enumeration district (ED) typically followed by a building identifier and dwelling number.

In situations of exceptionally difficult enumeration conditions there may be a small number of EDs where enumeration including dwelling listing has not taken place. **Every effort needs to be made to minimize this situation.** Adjusting for such nonresponse requires strong assumptions.

For each ED where data collection has taken place there will be a list of dwellings providing addresses and identifiers as described above, often in the form of a visitation record (VR). A code is provided for the dwelling occupancy status. There can be nonresponse to this variable, usually at a low rate, in cases where the enumerator is unable to determine the occupancy status of a dwelling. In some cases the VR also collects some basic information on household size and composition for occupied dwellings. There also can be nonresponse to some or all of these variables.

Questionnaire structures are quite similar with, in order

- Items for dwelling identification and tracking of enumerator activity
- Questions the enumerator can complete by observation, sometimes including dwelling occupancy status
- Household questions
- Individual questions, which get repeated for each usual resident

In most cases the household questions start with a list or roster of the usual residents, often also obtaining the sex and age/dob for each person. In one case we have seen these are the last questions in the household section. In another, this listing is preceded by a question re the number of usual residents. More commonly, a question(s) appears after this household listing in

which the number of usual residents is asked. Or there is no such question and the number of usual residents must be derived from the number of names in the list.

Now for some, perhaps even most, of the non-responding dwellings classified as occupied on Census Day we may have some additional information, often on the VR, such as:

- Number of persons
- Number of persons by sex
- Number of persons by sex and (approximate) age or age group (e.g. children under some specific age)

Thus nonresponse occurs at each of a number of stages:

1. EDs where data collection including listing did not take place (again, it is very important that this be minimized)
2. Listed dwellings where occupancy status is not determined
3. Occupied dwellings with nonresponse to demographic data to be collected on the VR
4. Occupied dwellings with nonresponse to the questionnaire

The question then is what should or can be done to account for the non-responding dwellings. Essentially there are three options available:

1. Do nothing
2. Adjust by weighting
3. Adjust by whole household imputation

1. Do Nothing

This is certainly the easiest approach. Conceptually at least, the nonresponding dwellings classified as occupied remain on the census database, identified as nonrespondents, and including the limited data referred to above that may have been available from the Visitation Record or other such control record. It means that tables of population counts can include persons in responding dwellings as well as persons in non-responding dwellings where the count of persons was available. Persons in other non-responding dwellings will be excluded and so population counts will be biased downwards to the extent of such nonresponse. Depending on the approach taken by the statistical office, tables of population by sex or age group or both will have counts that differ from that due to the differing amounts of available information for the non-responding dwellings. Tables for other variables will necessarily have to exclude all of the non-responding dwellings and so will be based on a reduced population. (Data quality notes would need to note the frequency of nonresponse and its exclusion from tables.) All of this produces an awkward situation for data users.

For statistical inference also this is not a good situation. First, to the extent there is no information on the nonresponding dwellings, the population counts will be understated. This is

not good and can be even more problematic if nonresponse rates differ by geography or any other characteristic of interest. Then as it concerns all the other data elements, the easiest and most natural assumption for users will be that the non-respondents have the same distribution as the respondents. But some users may make other assumptions that might lead to differing conclusions. Experience from past censuses tells us that the non-respondents are clearly not distributed the same as respondents. For example, in Canada, non-responding dwellings tend to have fewer persons than average; one person households are particularly over-represented in non-respondents. Further, these households tend to be younger than average. Last, males are over-represented in the non-responding dwellings. Consequently, **statistical inferences risk being misleading if the non-response rate is not small**.

Do nothing is not usually to be recommended.

2. **Adjust by weighting**

Weighting is a standard procedure for sample surveys. A variety of procedures are available to enhance the design weights by approaches such as post-stratification, calibration to known totals (often demographic projections), etc.

Sample surveys typically include an adjustment to the survey weights to account for unit non-response. Such adjustments are normally done within design strata (or sets of design strata) where sometimes quite a bit is already known about the population.

A standard census needs no adjustment for sampling of course. But there are non-response and other non-sampling errors, which as we noted earlier are not at random. For non-response adjustment, weighting assumes that within weighting class the non-response is at random. In a census context, can weighting classes be created with approximately those properties? Controlling for geography will usually help. Weighting classes will usually be designed to consist of sets of several geographically contiguous EDs. Depending on requirements and homogeneity of EDs and characteristics this might be as small as a supervisor district or as large as a parish.

Should there be any EDs where data collection did not take place, by a suitable weighting adjustment these can be represented by the EDs where collection was done. This requires an assumption that within each weighting group, EDs are reasonably homogeneous and that the "missed" EDs were essentially at random.

In a second step, weighting can be used to adjust for missing dwelling occupancy information. (Weighting group by weighting group, those with such information will represent those without.)

Third, if the VR includes variables for hhld size and composition, a similar weighting adjustment can be done to account for the cases where these variables are missing – occupied dwellings with such data will represent those for which those variables were not provided.

And in a last stage, responding dwellings can be weighted to represent the nonresponding dwellings. When available this basic demographic information on the VR can be used to improve adjustment for dwellings with no questionnaire by using it for construction of more homogeneous weighting classes for use at this last stage.

From a user perspective, weights are something of a complication relative to what might be the expected situation for many census data users. However, if this can be done such that it is reasonably transparent and easy for users through REDATAM or other census analysis and dissemination tool then it can be very effective.

Weighting is a clear improvement over doing nothing.

3. Whole Household Imputation

This section deals primarily with nonresponse to the questionnaire (i.e. the fourth stage noted above). But first we must consider nonresponse at earlier stages in the process.

When there are EDs where data collection was not done and there is no VR, there seems little choice but to adjust by weighting and have the ED represented by other EDs in the area. This makes the strong assumption that the "missing" EDs are similar to other EDs in the weighting area.

Imputation adjustment approaches are feasible at the other stages.

First, dwellings with unknown occupancy status. A reasonable approach is to impute an appropriate fraction as occupied based upon the fraction occupied amongst those with an occupancy classification in the same ED.

For treatment of nonresponse via imputation to any household size information on the VR, it can be treated as if it were part of the questionnaire. (This simplification is not possible in a weighting strategy.) As part of the initial data cleaning phase, it should be ensured that household size information on the VR is consistent with that available or derived from the questionnaire itself.

In this strategy for each non-responding dwelling (i.e. no questionnaire) classified as occupied on Census Day we search for a good quality donor dwelling based on the available known information for the non-responding dwelling. We will be able to use all of this information without discarding any of it. When a suitable donor is found, all of its data (except the names) is imputed to the non-responding dwelling.

Specifically, for each non-responding dwelling a search is made for possible donor dwellings that match on the available demographic information (for example, see the bullets on page one). Matching dwellings that are spatially close to the non-responding dwelling are usually preferred over those that are far away. This is often implemented by identification of the five (say) nearest matching neighbours and selecting one of these at random. In addition to using the matching variables, this stochastic element helps preserve the distributions. Use of a spatially near neighbour is based on the assumption that after accounting for the known matching variables, the composition and attributes of a near neighbour are more likely to be similar to those of the non-responding dwelling than a more distant donor record.

An additional constraint might be to prefer slightly more distant donors from the same enumeration district over closer donors from a differing enumeration district. This could help in situations where neighbouring EDs are distinctly different economically, say.

With the availability of GPS points for every building, computation of distances between dwellings should be a simple matter.

Alternatively, and more simply, it would be reasonable to select a donor at random from all matching records within the same enumeration district as the non-responding dwelling in question. (If no matching donor can be found in the same ED, one could then relax constraints on the quality of the match or search within neighbouring EDs.)

So, in increasing order of priority, a reasonable set of matching variables to use in a WHI procedure would be:

- Enumeration district
- Supervisor district
- Parish
- Household size and composition variables

The search for a suitable donor would start using all of these variables and a match found using as many of them as possible. As needed, variables would be dropped in order – first ED, then SD, then parish, then the hhld variables – until a match could be found.

It is important in this strategy also to keep track of how often each dwelling is used as a donor. The number of uses should be capped at some maximum limit. This could be set as a parameter

and initially set to a low value such as one. If this results in excess relaxing of matching constraints then a higher limit might be tried in areas where it results in problems.

An advantage over weighting is that imputation also allows to keep the total number of dwellings in any geographical unit consistent with the initial frame rather than only for geographical weighting classes.

Our opinion is that while weighting is a very good acceptable approach for non-response adjustment in a census, whole household imputation, while more complex to implement, can do a better job of preserving distributions and accounting for all of the information that is known for the non-responding dwellings.