Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Requirement: Collect less.
Our mission: Do the best we can.

O. Haag, P.-A. Pendoli, S. Faivre[1]

Abstract

In France, budget restrictions are making it more difficult to hire casual interviewers to deal with collection problems. As a result, it has become necessary to adhere to a predetermined annual work quota. For surveys of the National Institute of Statistics and Economic Studies (INSEE), which use a master sample, problems arise when an interviewer is on extended leave throughout the entire collection period of a survey. When that occurs, an area may cease to be covered by the survey, and this effectively generates a bias.
In response to this new problem, we have implemented two methods, depending on when the problem is identified:

- If an area is ‘abandoned’ before or at the very beginning of collection, we carry out a ‘sub-allocation’ procedure. The procedure involves interviewing a minimum number of households in each collection area at the expense of other areas in which no collection problems have been identified. The idea is to minimize the dispersion of weights while meeting collection targets.
- If an area is ‘abandoned’ during collection, we prioritize the remaining surveys. Prioritization is based on a representativeness indicator (R indicator) that measures the degree of similarity between a sample and the base population. The goal of this prioritization process during collection is to get as close as possible to equal response probability for respondents. The R indicator is based on the dispersion of the estimated response probabilities of the sampled households, and it is composed of partial R indicators that measure representativeness variable by variable. These R indicators are tools that we can use to analyze collection by isolating underrepresented population groups. We can increase collection efforts for groups that have been identified beforehand.

In the oral presentation, we covered these two points concisely. By contrast, this paper deals exclusively with the first point: sub-allocation. Prioritization is being implemented for the first time at INSEE for the assets survey, and it will be covered in a specific paper by A. Rebecq.

Keywords: Collection, sampling, R indicators

1. Review of household survey sampling at INSEE (Christine and Faivre 2009)

It is common practice in survey statistics to use two-stage sampling for face-to-face surveys. The data are collected by interviewers who spend some of their work time travelling between their homes and the residences of the households or individuals to be surveyed. The value of two-stage sampling is that it restricts the sample of households or individuals to a limited number of small geographic areas,

- to reduce collection costs (travel and search time) for the survey, which has a tight budget
- to allow for hiring a fixed contingent of interviewers, who receive training in survey methods and acquire experience in the area, both of which are beneficial in the long run.

For this method of survey sampling, we must

- define and construct the geographic areas (referred to as primary units)
- select the areas where households are to be chosen to get a representative sample of primary units (first stage of sampling)
- select the households or individuals, i.e., the secondary units, in those areas (second stage of sampling).

[1] National Institute of Statistics and Economic Studies (INSEE), France.

1.1 Interviewer activity areas, the primary units of INSEE’s household surveys

The construction of interviewer activity areas (ZAEs) is based on the methodology of the new census. In particular, the census distinguishes between small and large communes:

- A large commune has a population of at least 10,000. Part of it (8% of the dwellings) is surveyed every population census year, over a five-year population census cycle. **A large commune itself constitutes a ZAE.**

- A small commune has a population of less than 10,000. Each small commune is assigned to a particular census year, over a five-year population census cycle.

A small-commune ZAE is a group of small communes that satisfies the following criteria:

1. It contains at least one commune from every census year.
2. It contains at least 300 principal residences from every census year.
3. It belongs to just one administrative region (there are 22 in metropolitan France).
4. It has a radius of 20 kilometres or less.

1.2 First stage of sampling: Construction of the master sample

From the 3,832 ZAEs that make up the metropolitan territory, 567 are selected at random (Guggemos 2009).

1.3 Second stage of sampling: Dwellings

Depending on the subject of the survey, the statistical unit may be the household or the individual. In both cases, the statistical unit is reached through its dwelling.

The key principles are as follows:

- selection of the dwellings in the ZAEs of the master sample (these selected dwellings are commonly referred to as ‘address records’ for the interviewers)
- calculation of theoretical allocations for each ZAE with the aim of minimizing the dispersion of the sampling weights of the dwellings
- possible definition of upper and lower survey limits for each ZAE to take interviewer workload into account.

2. New work conditions for interviewers

The implementation of new work conditions for interviewers in 2013 means INSEE must match the annual collection time of an interviewer to a predetermined annual amount of work time. Collection time includes the following:

- total travel time, which depends on
  1. the number of dwellings to be surveyed
  2. the time required to get to the survey area (including the time to find the address)
  3. the average number of contacts required to make an appointment
- total survey administration time, which depends on
  1. the number of households to be surveyed
  2. the response rate
  3. the time to complete a questionnaire

As a result, INSEE carries out a planning process every year to assign all the samples for the following year’s surveys to interviewers in accordance with their available work time.

However, if an interviewer subsequently goes on extended leave during a collection period, there may be very few or even no interviews conducted for a given survey and a given ZAE. We refer to such ZAEs as ‘orphan’ ZAEs.

There is some leeway for making up the interviews in an orphan ZAE. There are two main options:

- make use of a reserve included in the work time of other interviewers in the system, intended for replacements[2]
- hire a casual interviewer (subject to the availability of space in the employment cap).

In some cases, however, after exploring these options, INSEE may not be able to free up additional collection resources to handle all or some of the interviews in the orphan ZAE.

Since a complete shutdown of surveying in a primary unit must be avoided, because it would create a bias[3] that cannot be corrected with the usual post-collection processing, a minimum number of interviews must be conducted in each ZAE in the master sample.

In this case, INSEE takes the following approach:

- carry out some of the interviews initially planned in the orphan ZAE; they must be assigned to other interviewers in the system, whom we will refer to as ‘alternate interviewers’
- relieve these alternate interviewers of some of the interviews that they were originally supposed to conduct in their home ZAEs, so that their total work time for this survey remains unchanged.[4]

This raises the question of what method should be used to select the dwellings that are surveyed, both in the orphan ZAE and in the home ZAEs of alternate interviewers.

Consequently, we undertook a project to devise a method of minimizing the loss of survey quality related to the decrease in field interviews, subject to the constraint of surveying a minimum number of households in each ZAE.

There are two possible scenarios:

- An interviewer is unavailable before collection or at the beginning of collection for a survey, and through the entire collection period. The method used in this case is sub-allocation, as described in this paper.
- The interviewer suddenly ceases collection work during a survey when much of the work has already been done, or the interviewer does not return during collection as planned. A different method must be used to prioritize the interviews of certain types of households.

3. Use of the sub-allocation technique

3.1 Context

Before collection or at the beginning of collection for a survey, we learn that an interviewer will be on extended leave through the entire collection period, orphaning the ZAE.

3.2 Formalizing the problem to be solved (Rebecq 2014)

Optimization criteria

1. Minimize the dispersion of the survey weights of the dwellings in the final sample (i.e., the sample that will ultimately be collected) in the region concerned.

[2] In 2014, initial planning allotted only 95% of the work quota of interviewers. The remaining 5% was held in reserve.

[3] Dwellings in the frame with zero probability of being included in the sample of respondents.

[4] Of course, since the travel time of interviewers to the orphan ZAE will usually be greater than the travel time to their home ZAE, we have to cancel more interviews in their home ZAE than we assign to them in the orphan ZAE.

2. Reduce the sample size by as little as possible.

Constraints

We applied three constraints in the model. They are described below. While the first constraint follows directly from the need to stay within the work-time cap, we added the other two to limit the impact that changing the sample has on the work conditions of the interviewers concerned.

1. The work time of the interviewers assigned to this operation will be unchanged.
   Assumptions about the work time of interviewers include the following, as described above:
   - **Travel time**. We assume that travel time depends solely on the distance between the interviewer's home and the principal commune in the ZAE being surveyed. The implicit assumption is that the average number of contacts per household is the same in the alternate interviewer's home ZAE as in the orphan ZAE, and that search times are the same, even though the interviewer is less familiar with the orphan ZAE. This assumption is therefore somewhat optimistic in terms of work quota consumed.
   - **Survey administration time**. We assume that this remains constant. This assumption is pessimistic in that, as we have seen, the interviewer will ultimately be assigned fewer interviews than originally planned. All things being equal, then, the interviewer will conduct fewer interviews, and his or her survey administration time will be shorter than originally planned.

   With these assumptions, the constraint only involves conserving total travel time.
2. Ultimately, an alternate interviewer will not be responsible for more interviews in the orphan ZAE than in his or her home ZAE.
3. Ultimately, an interviewer will remain responsible for at least 50% of the interviews initially assigned to him or her in his or her home ZAE.

Information needed to apply the method

- We must know which ZAEs are orphaned before collection or at the beginning of collection.
- We must know which alternate interviewers can be brought in.
- We must know how much work time is allocated to the survey for each of these interviewers.
- We must know the travel time between the homes of the alternate interviewers and the principal commune in the orphan ZAE, and between their homes and the principal commune in their usual ZAE.

Formalization of the problem

The region-by-region minimization program

$$(1) \quad \underset{n_{ZAE}}{Min} \frac{1}{n_{reg}} * \sum_{ZAE \in reg} n_{ZAE} * (Poids_{ZAE}^{\log} - Poids_{moy}^{\log})^2$$

$$où \qquad n_{reg} = n_{ZAE}^{orpheline} + \sum_{ZAE \in remplaçantes} n_{ZAE} + \sum_{ZAE \in reste\ reg} n_{ZAE}$$

$$Poids_{ZAE}^{\log} = \omega_{ZAE} * \frac{N_{ZAE}}{n_{ZAE}}$$

Under the constraints for each interviewer

$$(2) \quad n_{ZAE\,rempl}^{initial} * T_{ZAE\,rempl}^{enq} = n_{ZAE\,rempl}^{final} * T_{ZAE\,rempl}^{enq} + n_{ZAE\,orpheline}^{final} * T_{ZAE\,orpheline}^{enq}$$

$$(3) \quad n_{ZAE\,rempl}^{final} \geq \frac{n_{ZAE\,rempl}^{initial}}{2}$$

$$(4) \quad n_{ZAE\,rempl}^{final} \geq n_{ZAE\,orpheline}^{final}$$

3.3 Practical implementation

The best two or three scenarios (the optimal outcome of the program and the scenarios that come closest to it) are offered to the collection managers for the region concerned. They choose the scenario that best addresses the collection constraints.

Note that the ratio of the time it takes to reach the home ZAE to the time it takes to reach the orphan ZAE plays a more important role in the model than the travel time to the orphan ZAE. The larger this ratio is, the larger the number of interviews lost.

This justifies the introduction of special communications for the interviewers. The collection teams must validate the feasibility of each proposed scenario in the field.

3.4 Illustration with a fictitious example: Collection in the Vizille ZAE for the 2013 housing survey

Information available in the sample:

- The Vizille ZAE has an allocation of 32 households selected for the 2013 housing survey.
- The Rhône-Alpes region has a total allocation of 2,591 households for this survey.
- The dispersion of the sampling weights in the sample for this region is 1,954.
- The dispersion of the final sampling weights in the region if only one interview were to be conducted in the orphan ZAE would be 102,039.

Table 3.4.1
Information provided by the Rhône-Alpes regional directorate, which is responsible for overseeing collection in the Vizille ZAE: List of interviewers who can be brought in and their travel times

ZAE for which the interviewer is responsible	Initial allocation in the ZAE	Travel time of the interviewer to the usual ZAE (in minutes)	Estimated travel time of the interviewer to the Vizille ZAE (in minutes)	Ratio of travel times
La Mure	52	10	25	2.5
Saint-Marcellin	43	29	55	1.9
Grenoble 1	45	27	47	1.7
Grenoble 2	24	16	36	2.25
Grenoble 3	10	10	25	2.5

Interpretation: In the initial sample, there were 52 interviews to be conducted in the La Mure ZAE. The interviewer responsible for the La Mure ZAE takes an average of 10 minutes to reach the interview location. However, it takes him or her an average of 25 minutes (2.5 times longer) to conduct a survey in the commune of Vizille.

Table 3.4.2
Results obtained according to which interviewers are brought in

ZAE(s) for which alternate interviewer(s) are originally responsible	Subsequent sample size in Rhône-Alpes	Subsequent dispersion of the sampling weights
La Mure / Saint-Marcellin	2,544	1,955
La Mure / Saint-Marcellin / Grenoble	2,543	1,992
La Mure / Grenoble	2,540	2,059
La Mure	2,544	2,233
Saint-Marcellin	2,550	2,192
Grenoble	2,546	2,184
No alternate interviewer brought in (assumption: one address record completed in the orphan ZAE)	2,560	102,039

Interpretation: If we brought in only interviewers from La Mure and Saint-Marcellin to conduct interviews in the orphan ZAE of Vizille, the final sample size collected would be 2,544, and the sub-allocation would lead to a weight dispersion of 1,955. This dispersion is our quality indicator: the smaller it is, the better the sub-allocation. If we completed only one address record in the orphan area, the sample size collected would be 2,560 (the maximum conceivable value), but the weight dispersion would be the highest in the region, 102,039.

Details of the La Mure / Saint-Marcellin scenario (the best possible result)

In this scenario, we bring in the interviewers who are usually responsible for the La Mure and Saint-Marcellin ZAEs to do some of the interviews in the orphan ZAE of Vizille. To compensate, we relieve these two alternate interviewers of some of the interviews originally assigned to them in their home ZAEs to keep their work time constant.

The algorithm yields the following sub-allocations:

- The interviewers complete 15 of 32 address records in the Vizille ZAE.
- The interviewers complete 35 of 52 address records in the La Mure ZAE.
- The interviewers complete 30 of 43 address records in the Saint-Marcellin ZAE.
- The Grenoble interviewers are not brought in. The allocations in the three Grenoble ZAEs are unchanged.

In each ZAE (the orphan ZAE and the ZAEs of the alternate interviewers), the households that are eventually interviewed are selected at random from the households originally selected in the sample.

Hence, this sub-allocation 'costs' 15 address records, since we complete 15 of the 32 address records in the orphan area, but we lose 30 address records in the neighbouring ZAEs.

3.5 Results of the implementation

INSEE used the sub-allocation method for two of its household surveys:

- In the 2013 dwelling survey, six ZAEs (189 dwellings) were orphaned. Eight alternate interviewers were brought in. They surveyed 95 dwellings in the orphan ZAEs, and they were relieved of 138 interviews in their home ZAEs. Therefore, the cost of the method was a loss of 43 additional dwellings, though we avoided having an entire ZAE orphaned.

- In the 2014 survey of the living environment and security, six ZAEs (212 dwellings) were orphaned. Ten alternate interviewers were brought in. They surveyed 97 dwellings in the orphan ZAEs, and they were relieved of 176 interviews in their home ZAEs. Therefore, the cost of the method was a loss of 79 additional dwellings, though we avoided having an entire ZAE orphaned.

The method is currently being used for three other surveys:

- It is being used for the survey of household living conditions. This is a nine-year panel. Since interviewing for waves 6 to 9 is no longer mandatory, the interviews to be dropped in the alternate ZAEs were cancelled mainly from these waves. By contrast, all wave 1 interviews in the orphan ZAEs were completed.
- It is being used for the assets survey.
- It is being used for a survey measuring the resources of young people.

4. Conclusion

The sub-allocation collection adjustment method is central to the dilemma between bias and variance. It focuses on reducing bias, but it is accompanied by an increase in the variance of the results obtained.

It is designed primarily to avoid coverage biases in collection, which are difficult to correct, and to promote calibration-type post-collection processing by reducing the dispersion of the weights of responding individuals as much as possible before the total non-response processing stage.

However, it ultimately leads to a reduction of the number of respondents and, therefore, an increase in the variance of the results. This reduction is caused by the additional distance interviewers must travel to survey the dwellings in the orphan ZAEs in the event of sub-allocation. As a general rule, the number of surveys conducted in the orphan ZAEs is smaller than the number of surveys cancelled in the usual ZAEs of the alternate interviewers. In addition, since the orphan ZAEs are less familiar, search times increase. Lastly, since these ZAEs are farther away, there is sometimes less collection effort. Measures are under way to compare the response rates achieved by the same interviewer in his or her usual collection area and in the orphan area.

Thus, this method is not perfect, but it attempts in every way possible to reduce the loss of survey quality caused by a decrease in the sample size during collection.

References

Christine, M., and S. Faivre. 2009. “Le projet OCTOPUSSE de nouvel Échantillon-Maître de l’Insee.” In *Actes des journées de méthodologie statistique*, March 23 to 25, 2009. INSEE.

Guggemos, F. 2009. “Simulation de tirages de zones d’action enquêteurs pour les enquêtes-ménages de l’Insee.” In *Actes des journées de méthodologie statistique*, March 23 to 25, 2009. INSEE.

Rebecq, A. 2014. *Heuristique Branch-and-bound pour la sous-allocation et la réallocation*. Paper presented at the 8th colloque francophone sur les sondages, Dijon, November 18 to 20, 2014.