Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Moving from a census to tax sources: changing the survey frame to better coordinate INSEE samples

Thomas Merly-Alpa, and Ludovic Vincent[1]

Abstract

The Institut National de la Statistique et des Études Économiques (INSEE) has developed an alternative survey frame using tax information adjusted for administrative management rules. This change has led to the modification of the variables available for selecting and collecting samples; at the same time, the use of more advanced sampling methods such as spatial balancing for the selection of collection areas improves the accuracy of the institute's surveys. Using a single frame allowed for a coordinated selection of the Continuous Labour Force Survey (EEC) sample and INSEE's master sample, thereby facilitating the work of the interviewers.

Key words: Sample; Administrative base; Master sample; Spatial balancing; Coordination of samples; Indirect sampling.

1. Introduction

As a result of the improved availability and quality of administrative sources from taxation authorities, the Institut National de la Statistique et des Études Économiques (INSEE) has developed an alternative survey frame: the Housing and Individual Demographic File (Fidéli), which includes tax information that has been de-duplicated and adjusted for administrative management rules.

Using Fidéli instead of the French Census of Population, the survey frame traditionally used at INSEE, will reduce the size of survey areas and improve the sampling design. However as a result of this change information is lost, concepts are modified, and new variables are created to identify surveys that will be based on more precise geolocation data.

The renewal of the master sample, planned for 2020 as part of the Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes (Nautile) project and the Continuous Labour Force Survey (EEC) sample renewal, is based on this file which outlines the characteristics of a good survey frame, in particular completeness.

The goal in coordinating these two samples is to limit interviewer travel. The primary units in the master sample are drawn through spatially balanced sampling. The EEC sample is a set of compact clusters of approximately 20 dwellings also selected through spatially balanced sampling on proxy variables based on the labour market situation. They are coordinated by introducing coordination units (CUs) selected through indirect sampling via primary units, with the EEC clusters selected from the CUs in a final phase.

2. Using tax sources as a survey frame

2.1 INSEE survey collection

[1] Thomas Merly-Alpa, INSEE, 88 avenue Verdier Montrouge, France, 92120 (thomas.merly-alpa@insee.fr); Ludovic Vincent, INSEE, 88 avenue Verdier Montrouge, France, 92120 (ludovic.vincent@insee.fr)

Every year, INSEE conducts a number of household surveys (on rent, living conditions, etc.). A key characteristic of these surveys is that many of them are conducted in-person and require an interviewer to be present. For these interviews, INSEE implemented the following method:

- Select a sample from the collection areas (the primary units) within an acceptable range that is representative of the entire French territory.
- Within each of the selected collection areas and for each survey, sample the dwellings to be interviewed.

This area sample is known as the master sample.

The current master sample, introduced by Faivre and Christine (2009), had a ten year life-span. There were many constraints for the areas selected. They were to:

- Contain dwellings from each of the rotation groups in the Census of Population.
- Located within as narrow as possible geographical range in order to minimize interviewer travel.
- The number of dwellings had to be large enough to avoid surveying the same dwelling more than once within a five year period.

The current system, described in Godinot (2005), follows the Census of Population closely from a statistical perspective (such as the use of census variables for stratification) as well as for collection (use of the image of the completed address in the census bulletin).

2.2 A new source: Fidéli

INSEE has to renew the master sample for 2019. As a prelude to this process, and using the preliminary analyses conducted by Hallepée, Pendoli and Sautory (2018), INSEE began working on balancing the census rotation groups which are based on the results of the 1999 exhaustive census. It was discovered that sociodemographic trends differ from one group to another and increase each year. Therefore the imbalance across census rotation groups over time has a direct negative impact on the quality of the current master sample and of any master sample based on the same selection method.

Fidéli (the Housing and Individual Demographic File) is a new source that presents new opportunities for survey sampling and for the establishment of zones or the identification of the units to be surveyed.

Fidéli, presented in Lollivier (2015), is a file of individuals and dwellings from tax files that have been reconciled and completed by the communities and residential hotel directories. The have been enriched with geolocation information (addresses, zoning) and income information from Filosofi (Fichier Localisé Social et Fiscal). These different processes make it possible to create a file that has the characteristics of a good survey frame:

- <u>Completeness</u>: unlike a census the tax source covers the entire population, a significant advantage with regard to the accuracy of household surveys.
- <u>Uniqueness (no duplicates)</u>: INSEE's work on tax files makes it possible to clean data and ensure (as far as possible) the uniqueness of each individual and each dwelling. This facilitates the collection and the design of the sampling plan and also improves the accuracy of the surveys.
- <u>Timeliness of the data</u>: the uploading and processing of tax data by INSEE provide a complete survey frame within 18 months; an advantage over the census which is as fast as the last annual census and covering one-fifth of the population.

Additionally, the switch to Fidéli allows individual selection. Using tax files makes it possible to have identification variables and sufficient information on each individual to accurately identify the individuals of interest. INSEE therefore decided to draw new samples for future household surveys from Fidéli's official statistical system.

2.3 New variables for sampling and collection

Transitioning from the census to Fidéli is not limited to removal of the five rotation groups for a comprehensive frame. As with any source change, some variables disappear, others emerge, and some concepts are modified. It will no longer be possible, for example, to use data such as level of education and workers' socio-professional category directly as these data are not included in the tax files. It will be necessary to use a proxy or less detailed sources. In contrast it will be possible to use other variables such as the income details. It will also be possible to more accurately pinpoint the location of collection units using geolocation variables.

Finally, some variables (or concepts) are present in both sources; however, the definitions and quality may vary (e.g. variable pertaining to social housing or status of a residence – main or secondary). Overall, Fidéli contains many additional variables over the census, and the quality will improve every year.

These changes impact three different levels:

- Before the selection, stratified sampling will be conducted on different variables, closer to or further from the subjects depending on the surveys. In addition, the translation of "regular census dwelling" as a unit of interest, or more generally the surveys' field of enquiry will change from what currently exists.
- In the collection phase, identifying the units surveyed cannot be based on the data currently being used as these are taken from the census bulletins. If the census data are no longer available, Fidéli provides several addresses of varying quality that will assist in finding the dwelling. Additional information (mail, telephone numbers, geographic coordinates…) could be made available to improve unit identification; causing the associated method to evolve.
- Following the collection, the survey frame can be used to make adjustments to the data. Surveys using census variables to adjust for total non-response will be reviewed for adaptation to the new survey frame. The calibration variables are at aggregated levels (e.g. at the community level and not that of the sample unit) should not be modified. That said, Fidéli could provide new variables that can be useful for surveys in improving the calibration process.

3. Coordinating the master sample and the EEC

3.1 The Nautile master sample

The purpose of the master sample is to produce a first stage geographical sample selection in order to concentrate the collection of several surveys in the same areas. Sample selection for the surveys therefore occur in the second stage within the selected areas. It is necessary to first partition the territory into primary units. This method consists of grouping communities to create areas with a sufficiently large population, thereby ensuring that individuals are not interviewed more than once during the life-span of the master sample. These areas must nevertheless be small enough to limit travel time during the survey. This method is based on an algorithm solution for the travelling salesman problem presented by Applegate et al. (2003); identify the best possible path of each French department, and then paths are subdivided into primary units. Favre-Martinoz and Merly-Alpa (2017) present a more detailed version of this method which resulted in the construction of 5,128 primary units.

Once the master sample has been used to draw most of INSEE's household surveys, it seems pertinent to try to balance the selection across as wide a range of socioeconomic variables as possible. In addition, the preliminary research carried out by Favre-Martinoz and Merly-Alpa (2016) has revealed that spatially balanced sampling has significant advantages when drawing a master sample. By limiting the selection of geographically close units that share socioeconomic characteristics, this method improves the accuracy of variables especially those excluded from the balancing variables. It also results in decreased deterioration of the accuracy of variables over time.

Using the community as a building block of the primary units provides many useful variables from the French Census of Population and various administrative sources. We are, however, aware that using too many balancing variables can cause reduction in quality; the first simulations carried out with the entire set of balancing variables did not respect the fixed size constraints and thus there was variation in the number of survey areas.

Guillo (2018) has developed a way to reduce the number of balancing variables; synthesize the information in all of the variables by data analysis methods. Applying principal component analysis to the primary units of the data set resulted in 15 axes explaining 99% of cloud data inertia. Balancing on these axes makes it possible to have increased accuracy in estimating cloud variables as well as those which are correlated, while complying with the fixed size constraints.

3.2 The Continuous Labour Force Survey

The Continuous Labour Force Survey (EEC) is a survey aimed at examining the situation of people in the labour market both structurally and cyclically. It is the French equivalent of the Labour Force Survey (LFS). In France, it is the only source that provides a measure of the concepts of activity, unemployment, employment and inactivity as defined by the International Labour Organization. This is an areolar survey: the sample is a set of geographically compact areas known as clusters that are grouped into sectors. A cluster is a group of approximately 20 dwellings; there are six or seven clusters in each sector. Every quarter, comprehensive interviews are conducted with one cluster from each sector over a two-week period; after six interviews, the process begins with the next cluster in the same sector. This interview mode, developed by Loonis (2009) for a period of nine years, will be used for the future EEC sample.

Unlike what was presented earlier in this paper, the EEC sample is not intended to represent a broad range of subjects. The EEC focuses on labour market topics. It is necessary to be specific solely on these topics, with potential variations according to the dissemination areas. There are two possible improvements for this purpose. Firstly, as discussed above, we use spatially balanced sampling to select the 2944 sample sectors. On the other hand, the variables at the communal level are replaced by *proxy* variables of the survey concepts. These variables at the individual or housing level are constructed using frame information to get as close as possible to the concept of interest. We use the perception of incentives for job seekers as a proxy for unemployed status, even though the two situations are clearly not equivalent. The choice of variables is detailed by Costa (2018).

3.3 Coordinating the two samples

Drawing the master sample on the one hand, and the EEC sample on the other hand, involves selecting the geographical areas in which the interviewers will conduct the interviews. The mobilized samples were not drawn at the same time as the distance between an EEC sector and a selected primary unit does not matter. The only factor taken into account is the separation; in order to avoid re-interviewing the same households, the EEC sample is removed from the selected primary units it intersects.

This independent management poses several problems. First, it involves fairly long trips when an interviewer has to travel to remote regions. Moreover, the geographic dispersion of the collection areas reduces the possibility of replacing interviewers in the event of prolonged absence (i.e. illness, etc.).

One solution is to concentrate the collection in such a way so that the sectors and the primary units selected are close; however, this approach results in a cluster effect which may adversely impact the accuracy of the survey estimates. Matei (2016) describes several possible methods for obtaining good quality data.

The goal is to establish wider areas that include the primary units and to consider the coordination process as draws within the areas; called coordination units, primary units and of EEC sectors. There are two possible alternatives. Using the direct method, one can sample the coordination units, followed by the primary units and the sectors within the selected coordination units. Alternatively, one can sample the primary units, deduct the selected coordination units and draw the sectors from the coordination units. This is the indirect method, referring to the indirect sampling of coordination units via primary units, a concept introduced by Deville and Lavallée (2006).

Several arguments and results lead to favor the indirect method. First, it seems difficult to balance the primary units with the constraint of selecting a primary unit within each coordination unit. If selected coordination units are restricted from containing selected primary units, the benefits of coordination are reduced. However, simulations show that with regard to the accuracy of many variables in the master sample, the indirect method is better than the direct method.

Nevertheless, the indirect method needs improvements. Indeed, the indirect drawing of the coordination units does not guarantee the quality of the coordination units sampled. As described earlier, drawing sectors of the EEC is balanced on specific variables. In order to ensure good quality balancing, the sample universe (the collection of all of the coordination units selected by indirect sampling) must have characteristics similar to those of the entire population. The weight share method defined by Deville and Lavallée (2006) may guarantee this, as each coordination unit can come into contact with *n* different primary units (corresponding to the number of links) and can be captured several times if several of these primary units are selected. Paliod (2018) proposes introducing variables transformed into so-called indirect variables that make it possible to balance the universe of selected coordination units. These variables are generated taking into account the number of links allowed to be in contact with a coordination unit.

4. Conclusion

The transition from census to tax sources to construct household survey frames is an opportunity for INSEE to renew and improve its sampling methods (spatial balancing, new proxy variables for the survey topics, coordination), organization of collection (grouping primary units and EEC sectors, new contact information), and post-processing (new variables, new calibration margins). This evolution is an organizational and methodological challenge, certain aspects of which are not yet understood. The stability of administrative tax sources is not certain: passage to the withholding tax starts in January 2019, the possible end of the housing tax…

References

Applegate, D., W. Cook, and A. Rohe (2003), “Chained Lin-Kernighan for large traveling salesman problems”, INFORMS Journal on Computing, 15(1), 82-92.

Costa, L., T. Merly-Alpa, and M. Chevalier. (2018), “Le renouvellement de l'échantillon Emploi : améliorations et évolutions”, Actes des Journées de Méthodologie Statistique de 2018, Insee.

Deville, J.-C., and P. Lavallée (2006), “Sondage indirect : les fondements de la méthode généralisée du partage des poids”, Techniques d’enquête, 32(2), p. 185.

Favre-Martinoz, C., and T. Merly-Alpa (2016), “Utilisation des Méthodes d'Échantillonnage Spatialement Équilibre pour le Tirage des Unités Primaires des Enquêtes Ménages de l'Insee”, 9eme Colloque Francophone sur les Sondages, Gatineau.

Favre-Martinoz, C., and T. Merly-Alpa (2017), “Constitution et tirage d’unités primaires pour des sondages en mobilisant de l’information spatiale”, 49èmes Journées de Statistique, Avignon.

Faivre, S., and M. Christine (2009), “Le projet OCTOPUSSE de nouvel Échantillon-Maître de l'Insee”, Actes des Journées de Méthodologie Statistique de 2009, Insee.

Godinot, A. (2005), “Pour comprendre le recensement de la population”, Insee Méthodes, hors série - mai 2005.

Guillo, C., and T. Merly-Alpa (2018), “Un nouvel Échantillon-Maître pour 2020 et pour Nautile”, Actes des Journées de Méthodologie Statistique de 2018, Insee.

Hallépée, S., P. A. Pendoli, and O. Sautory (2018), “La re-pondération des enquêtes annuelles de recensement pour une diffusion complémentaire du RP”, Actes des Journées de Méthodologie Statistique de 2018, Insee.

Lollivier S. (2015). *Le répertoire statistique des logements*, Commission Territoires du CNIS.

Loonis, V. (2009), “La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation”, Actes des Journées de Méthodologie Statistique de 2009, Insee.

Matei, A., and A. Grafström (2016), “Coordination des échantillons dans l'échantillonnage spatial”, 9eme Colloque Francophone sur les Sondages, Gatineau.

Paliod, N., M. Chevalier, and T. Deroyon (2018), “Coordination spatiale d'échantillons : application à l'EEC et l'Échantillon-Maître”, Actes des Journées de Méthodologie Statistique de 2018, Insee.