

The innovative sample design of the French New Annual Master Sample 2009 : building fresh annual sample frames for household surveys based on the new annual Census

Berlemont Bruno¹, Christine Marc², Faivre Sébastien³

¹INSEE, bruno.berlemont@insee.fr

²INSEE, marc.christine@insee.fr

³INSEE, sebastien.faivre@insee.fr

Abstract

For more than forty years, the sample frames for national household surveys carried out by INSEE (except the Labour Force Survey) have been built from the lists of dwellings established by the Census. A new rotative Census has been taking place in France since January 2004. Instead of an exhaustive counting of the whole population, it consists of yearly surveys on a part of the territory, based on samples of municipalities or addresses, a sample of them being covered by the Census each year during a five years cycle. This very new technique implies to redefine the sample designs of all household surveys. One of the most important principles is to use as a frame for the surveys of a current year the lists of dwellings covered by the Census of the previous year. But, since most of the surveys are face to face, the dwellings which must be surveyed have to be concentrated in some areas, in order to reduce the costs. It implies to build *primary units* (PU), among which a sample is drawn, and to employ a network of interviewers located not far from those PU. The automatic process of building of those PU is presented. A major methodological issue was then to establish the sample frame of the PU. Then the final step was to check if the annual sample frames formed by the censused parts of the drawn PU are representative of the whole country, leading to establish a solution of calibration of the PU.

Keywords: Master Sample, Rotative Census, Primary Units

1. Main innovation of the New Master Sample 2009: a fresh sample frame each year in a Master Sample system!

For more than forty years, the sample frames for national household surveys carried out by INSEE (except the Labour Force Survey) have been built from the lists of dwellings established by the Census. Until the New Master Sample starts in 2009, French household survey samples are drawn in the French Master Sample 1999 based on the Census of 1999.

A new rotative Census has been taking place in France since January 2004. Instead of an exhaustive counting of the whole population, it consists of yearly surveys on a part of the territory, based on samples of municipalities or addresses, a sample of them being covered by the Census each year during a five years cycle.

In big municipalities (10 000 inhabitants or more), a sample of addresses is drawn each year in order to have 8% of the cities' dwellings being censused:

- Building in each of them of 5 samples of *addresses* (« rotation groups ») *from a file updated each year (RIL, register of located blocks)*.
- Drawing each year a sample of dwellings (clusters of addresses) ; the average sample rate is about **40%** of all dwellings belonging to the current rotation group.
- Census of this sample of dwellings.

Small municipalities (less than 10 000 inhabitants) have been split in five rotative groups (random balanced sample of municipalities with equal probabilities), one of the five groups being censused each year (whole Census of all municipalities belonging to the group).

This very new technique implies to redefine the sample designs of all household surveys, the Census becoming the first phase of the new sampling design. It has led us to deal with very innovative methodological issues, with the building, sampling and calibration of a new kind of primary units.

This paper will present the main opportunities of this renovation to improve the quality of survey samples thanks to fresh annual sample frames, and the current methodological issues.

One of the most important principles is to use as a frame for the surveys of a current year the lists of dwellings covered by the Census of the previous year. **This principle will make data collection easier (less wastes), will avoid using a specific system for new dwellings and will be more convenient to select samples on given sub-populations.** But, since most of the surveys are face to face, the dwellings which must be surveyed have to be concentrated in some areas and not spread all over the territory, in order to reduce the costs. It implies to build *primary units* (PU), among which a sample is drawn, and to employ a network of interviewers located not far from those PU.

2. New Primary Units in order to combine a fixed network of interviewers and a new sample frame each year

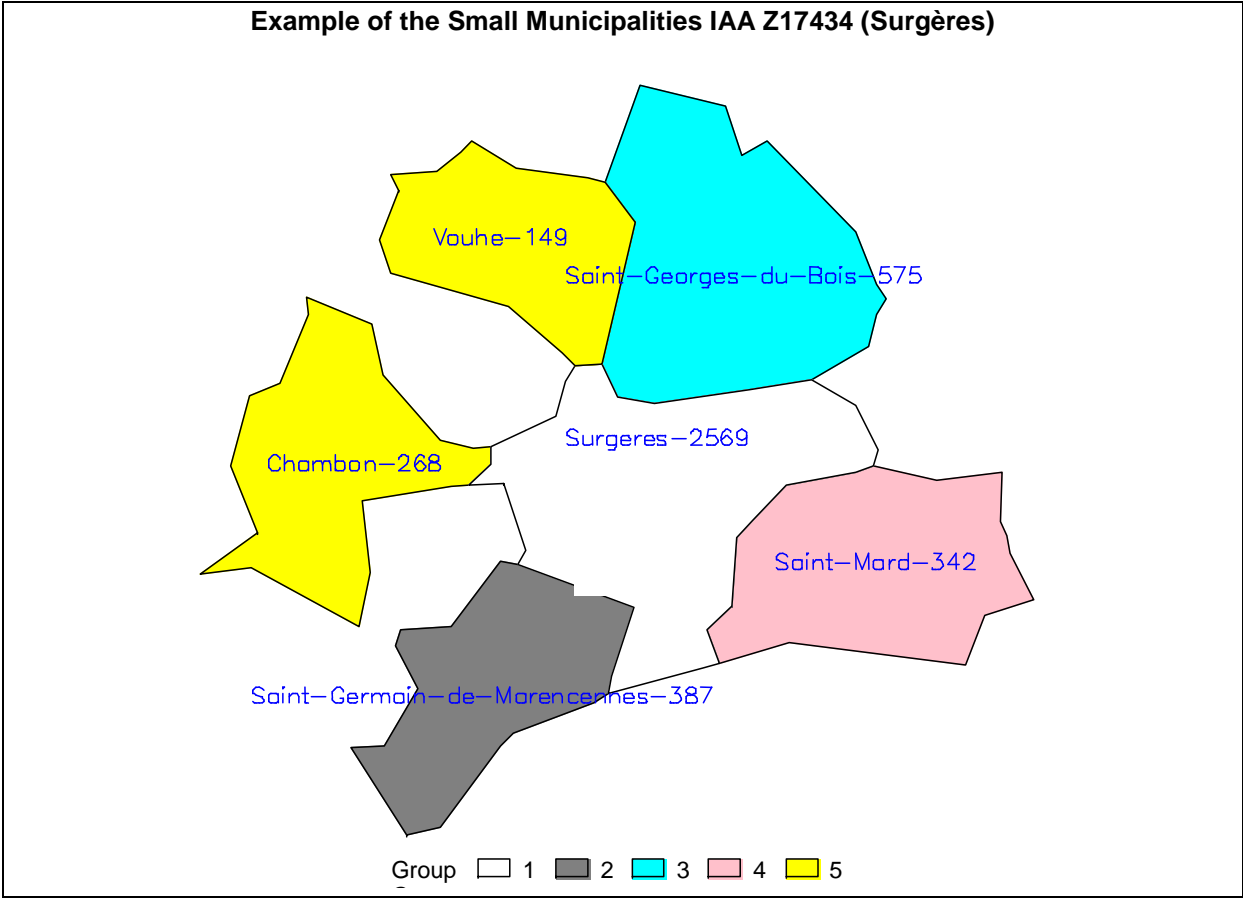
The first step was then to build *fixed* PU, under constraints of size and conditionally to the Census sample of addresses or municipalities. As for the minimum size of PU, studies on the current Master Sample had shown that up to 300 dwellings could be drawn yearly in households sample, which led to establish the principle that there should be a minimum of 300 dwellings in each PU (to be able to select 300 dwellings each year).

We focus here on some specific aspects of the constitution and selection of the Primary Units (IAA, Interviewer Action Area) of the French Master Sample. One must remember that the Primary Units are sampled at once and then used for all the dwellings samples drawn (which means that the sample frame used to draw the dwellings samples is built up with all the dwellings covered by the Census of the

previous year and located in the drawn Primary Units). The drawn IAA sample is meant to be used during ten years (between 2009 and 2019).

Due to the fact that more than 250 dwellings are censused each year in all big municipalities (and more than 300 dwellings in the great majority of them), it was decided that each big municipality would be itself an IAA. This led to build 850 Big Municipalities IAA.

The challenge was then to build IAA with the 35 000 small municipalities, taking into account the threshold of 300 dwellings in each Rotation Group that was established. Under this constraint, the challenge was to have an “optimal” IAA partition with Primary Unit as less extended as possible.



The Small Municipalities IAA shown as example is built of 6 small municipalities, in order to reach the threshold of 300 dwellings in each rotation group. The “center municipality” of the IAA is the municipality of Surgères, affected to the first rotation group (this concept of “center municipality is further explained in section 3). It can be noticed that the “center municipality” is the most important municipality of the IAA (2569 dwellings at the General Census of 1999) : it happens to be so for the great majority of the built Small Municipalities IAA. The other municipalities of the IAA are Saint-Germain-de-Marencennes (rotation group 2; 387 dwellings), Saint-Georges du Bois (rotation group 3; 575 dwellings), Saint-Mard (rotation group 4; 342 dwellings), Chambon (rotation group 5; 268 dwellings) and Vouhé (rotation group 5; 149

dwelling). It can be noticed that it was necessary to have two municipalities in rotation group 5 in order to reach the threshold of 300 dwellings in this rotation group.

In this case, the small municipalities belonging to the Small Municipalities IAA are contiguous (which minimizes the extend of the IAA and so the distances between the selected dwellings in the IAA), but it was not always possible to do so.

If the IAA is active (drawn in the IAA sample), the selection of dwellings in the IAA will be made as follows:

In January 2009, the first rotation group is censused. In the IAA, the municipality of Surgères is censused, which means that the “second stage sample frame” in the IAA is formed with dwellings censused in the municipality of Surgères. The lists of dwellings censused each year in January are available at the end of the year, which means that all survey samples drawn in 2010 are drawn in the 2009 Census dwellings lists: in the IAA, all dwellings of those samples are then drawn in the municipality of Surgères.

In January 2010, the second group is censused, corresponding in the IAA to the municipality of Saint-Germain de Marencennes. From January 2011 (dwelling lists from the 2010 Census available) to December 2011, all dwellings drawn in the IAA are drawn in the municipality of Saint-Germain-de-Marencennes.

Following the same principle, all dwellings drawn in the IAA in 2012 are drawn in Saint-Georges-du-Bois (rotation group 3, censused in January 2011), all dwellings drawn in the IAA in 2013 are drawn in Saint-Mard (rotation group 4, censused in January 2012) and all dwellings drawn in the IAA in 2014 are drawn in Chambon and in Vouhe (rotation group 5, censused in January 2013).

3. Specific algorithm established to build the Small Municipalities IAA

Due to the important number of small municipalities that had to be affected to an IAA (35 721 small municipalities) and the specific constraints imposed to the IAA (a minimum of 300 dwellings in each of the five Rotation Groups established by the French Census), an automatic process of building was set up.

The algorithm works as follows:

One IAA is build around a municipality called “center municipality”. Other municipalities are then tested to be allocated to the IAA, starting with the municipality closest to the center municipality and going on at each step with the closest municipality that has not been tested yet, in order to reach the threshold of 300 dwellings in each rotation group. Other municipalities can be allocated to the IAA only if:

- they belong to the same region that the “center municipalities”, as IAA are built within regional borders.

- they are still available (they have not been affected to another IAA built before)
- their distance to the “center municipality” does not exceed a fixed threshold (finally chosen to be 20 km, after having tested several values, see below)
- the threshold of 300 dwellings in their rotation group has not been reached yet

An IAA is achieved if, among municipalities of the same region (not yet allocated), whose distance to the pivot is less than a given threshold, it is possible to find enough municipalities in order to reach 300 main dwellings in each rotation group.

If not, the building of the IAA fails and all the tested municipalities remain available. Another municipality is then tested as “center municipality” to build an IAA.

In each French region, the algorithm works then as follows (“flying phase”):

- First, the most important small municipality of the region is tested as possible “center municipality” to build an IAA
- At each step, the biggest municipality not yet allocated to an IAA is tested as possible “center municipality”

The flying phase ends when the smallest available municipality has been tested as “center municipality” to build an IAA.

Remaining municipalities are then allocated to the closest IAA during the second phase of the algorithm, called “landing phase”, only if their distance to the “center municipality” of the closest IAA does not exceed the fixed threshold (same threshold as in the flying phase).

An important parameter of the algorithm was the maximal distance of the tested municipalities to the center municipality of the IAA. If this threshold was too low, very few IAA could be built during the flying phase, and many municipalities were unaffected after the landing phase, leading to affect remaining municipalities to an IAA located “far away” from the municipality and therefore to build extended IAA with many municipalities. But if the threshold was too high, it enabled to build very extended IAA during the flying phase and the landing phase, which also led to have extended IAA.

Several values of the parameter were tested, in order to find out the optimal threshold. Two criteria were used to find out the optimal threshold:

- the number of unallocated municipalities at the end of the “landing phase”
- the average extent of the built Small Municipalities IAA, calculated as the maximum (on the five years of the Census cycle) yearly distance between the “center municipality” of the IAA and the dwellings of the small municipalities of the IAA belonging to the censused rotation group.

Maximal distance to the IAA "center municipality"	Number of built Small Municipalities IAA	Number of small municipalities unallocated after the landing phase	Average extent of the built small municipalities IAA (in km)
10	1 788	10 996	7,8
15	2 565	1 746	10
18	2 779	645	10,9
19	2 848	465	11,2
20	2 886	363	11,4
21	2 944	247	11,7
22	2 969	175	11,9
23	3 005	130	12,1
24	3 037	107	12,3
25	3 056	83	12,5
26	3 093	68	12,7
27	3 115	32	12,9
28	3 144	15	13,2

Thanks to this results, it could be considered that 20 km was a reasonable value for the threshold parameter, as it led to a reduced number of small municipalities still available after the flying phase (363) without increasing too much the extent of the built IAA (maximum yearly average distance of a selected dwellings to the "center municipality" of 11,4 km).

Seven additional Small Municipalities IAA could be built with the remaining municipalities, which led to have finally 2893 Small Municipalities IAA with the 35 721 small municipalities that had to be affected to an IAA.

5. Sampling of the IAA

A major methodological issue was then to establish the sample design of the IAA sample. Basic hypotheses of the sample design are :

- IAA are drawn proportionally to their sizes (number of main dwellings)
- Some of them are systematically kept (« take-all strata »).
- Except for take-all stratum IAA (where several interviewers can be employed in the same IAA), one interviewer works in one IAA

The number of IAA to be drawn was then established with following parameters : For a common sample size with sampling rate 1/ 2000 (a little less than 12.000 main

dwelling), the average allocation is 20 sampled units for each interviewer and therefore for each IAA (except take-all stratum).

It was found out that the threshold of the take-all stratum was 40 000 dwellings (meaning that the 37 IAA with more than 40 000 dwellings belong to the take-all stratum and the 3706 remaining IAA belong to the sampling stratum), and that a sample of 488 IAA had to be drawn among the 3706 IAA of the sampling stratum.

The sample design of the IAA sample is stratified according to the 22 French **regions (except overseas territories)** and balanced on regional totals. The balanced sample algorithm CUBE developed by Deville and Tille was used to draw the final IAA sample: it ensures that the estimated total of auxiliary variables in the sample is exactly the real total in the population.

A major issue is that it appears to be necessary to have balancing conditions not only on the level of whole IAA **but also for each rotation group** in order to benefit each year from a « representative » sampling frame, which leads to increase the number of balancing conditions and reduce the number of allowed independent variables.

Different sample designs with different auxiliary variables (introduced at rotation group level for each IAA or at the whole IAA level) were compared through sample simulations in order to establish the best IAA sample design. For each studied sample design, 1000 samples of IAA were drawn: on each of those 1000 samples, the Horvitz-Thompson estimator of the total of a wide set of socioeconomic variables (data taken from the 1999 General Census or administrative data such as Tax data) was calculated. The variance of the 500 estimations was then calculated, in order to evaluate the quality of the sample design.

This work led to select the following balancing conditions:

- ***Number of main dwellings of municipalities belonging to the IAA, for each of the five rotation groups.***
- ***Total income (from tax sources) of municipalities belonging to the IAA, for each of the five rotation groups.***
- ***Total number of dwellings in the whole IAA in peri-urban areas, rural areas and urban areas.***

Additional balancing conditions could be introduced in region Ile-de-France (Paris region), thanks to the quite important number of IAA drawn in the region (84).

6. Calibration of the IAA

Then the final step was to check if the annual sample frames formed by the censused parts of the drawn IAA are representative of the whole country, comparing the estimate (from the sample of IAA) of totals of different auxiliary variables (the values of which are supposed known on whole IAA) with the true total in France (known through Census 1999 or other comprehensive data, such as tax sources).

Some results tended to show that some specific populations were underestimated by the drawn IAA (dwellings located in rural areas for example, as shown below) or overestimated (urban areas).

Rotation Group	Relative error of yearly sample frames on the variable “number of dwellings in rural area”
GR 1	+3,4%
GR 2	-3,3%
GR 3	-7,9%
GR 4	-8,1%
GR 5	-9,4%

A solution of calibration of the IAA was therefore established, in order to obtain yearly representative sample frames.

The final weights of the censused parts of IAA are then calculated each year by minimizing the distance to the initial sample weights under following conditions

$$\forall t \in \{1, \dots, 5\}: \sum_{k \in s} \omega_k T_{k,t}(Z) = T(Z)$$

Where $T_{k,t}(Z)$ = total of the calibration variable Z on municipalities belonging to rotation group t in IAA k, $T(Z)$ = true total of variable Z in whole population and ω_k = new weight of IAA k after calibration process

Auxiliary variables used for calibration are:

- *balancing variables*
- *number of employed people, according to the sector of activity*
- *number of dwellings according to the size of urban units (built-up areas).*

It was then checked that the calibration of the IAA leads to have a more representative Master Sample, by showing that relative error equals zero for all calibration variables and does not increase for other variables of interest.

7. Sampling of dwellings within the IAA

The sample frame in each IAA is then built up each year with the list of dwellings censused in the IAA during the last yearly Census Survey.

In small municipalities IAA, all censused dwellings are included in the sample frame.

In big municipalities IAA, a resampling of censused dwellings is carried out in order to take into account the fact that the sample rate of some specific addresses ("big addresses" with more than 60 dwellings or "new addresses" that have been built for less than five years) have a bigger probability to be censused than remaining addresses. Only a part of dwellings belonging to specific addresses are selected in the sample frame (with a systematic random sampling procedure), in order to have in the sample frame the same proportion of dwellings belonging to specific addresses that in the whole municipality.

The number of dwellings to be drawn in each IAA is then calculated in order to minimize the variations of the final weights of dwellings in the sample, under additional constraints of minimum and maximum number of dwellings drawn in each IAA to provide a reasonable amount of work to the interviewer in charge of the IAA.

In each IAA, dwellings are then drawn with equal probabilities with a systematic random sampling procedure.

The final sample weight of the sampled dwellings is then calculated taking into account all the sample phases: Census sampling and IAA sampling that lead to the Master Sample's yearly sample frame (with also the IAA calibration step) and dwellings sampling in the Master Sample's yearly sample frame.