United Nations ECE/CES/GE.41/2010/20

Economic and Social Council

Distr.: General
5 July 2010

English
Original: French

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Thirteenth Meeting
Geneva, 7-9 July 2010
Item 3 of the provisional agenda
Plans for census quality evaluation

Census quality

Note by the French National Institute of Statistics

Census quality

Michel CÉZARD, Olivier LEFEBVRE, INSEE

December 2008

Abstract
Since 2004, the French census system has worked to a revised method combining two core principles: a rolling census based on a 5-year cycle, and the deployment of a sample survey scheme led on communes *counting either side of 10,000 inhabitants, based on a comprehensive housing inventory called the 'RIL'. These principles are written into national census law. This revamped census system can output detailed year-in-year-out statistics at each geographic level of analysis, from national figures down to commune or district-level scope (§ 1 to 3).*
The census reform was implemented in response to a mounting need for recent demographic data, most importantly at more granular levels. The reforms also make it possible to even out both the HR and budget-related census costs, and implement better operational control (§ 4). Finally, the new census format, and particularly the housing inventory, cuts down on omissions.

The population census, in its French format, meets the five core criteria set by the UN (§ 5) as well as the general quality assurance criteria governing statistical operations (§ 6).

Quality is managed end-to-end throughout the census preparation, collection, processing and publishing operations. Census quality is assured through quality control and metrics protocols plus reviews with input from all the census actors involved: communes*, data collection agents, INSEE teams, and subcontractors (§ 7 and 8). Pilot tests and phased investigations are run to enhance a quantitative and qualitative analysis of census coverage (§ 9).*

A/ General overview of the census process

1. Legal foundations of the census

The census is founded on French law passed in 2002 setting out the purposes and core principles of the census: census organization and canvassing protocols, the framework governing INSEE-*communes* partnership, the principle of annually-published population figures for every single *commune*. This law is supplemented by several acts and decrees shaping how the census is to be applied.

The population figures drawn from the census serve as key reference data for the implementation of numerous policy provisions implemented at local community level (State pay-outs to local *communes*, organizational setups for local councils, status and pay schemes for local authority staff) as well as for community planning, gridding pharmacy coverage, and so on. In total, there are over 350 policy provisions all referenced to a population count, which therefore has to be delivered with exemplary quality for each of the 36,700 French *communes*, including close to 10,000 that count less than 200 inhabitants.

The French census is also compatible with the European regulations issued on 12/08/2008, and the census method employed is listed among the recognized methods for providing statutory information.

2. Method

The first principle of the method is to dispatch the data collection work over a rolling five-year rotation and to release information for each year based on the median year of the cycle. Consequently, data is generated for each year based on the data collected for the years A-4 to A and collated to represent year A-2.
The second principle of the census method is to survey the largest *communes*, where there is an adequately large population to enable the survey to generate robust data. This sampling scheme is designed to lighten the census burden on both sides, not just for the INSEE but also for the *communes* and the population canvassed, while also improving census collection quality and reducing the omissions rate, which is higher in big towns. The census survey hinges on the '*RIL*', which stands for *Répertoire d'Immeubles Localisés*, a comprehensive buildings inventory.

The sheer number and type-range of the French *communes* (close-on 37,000 *communes*, a quarter of which count less than 200 inhabitants) dictated the implementation of a specially designed data collection system.

***Communes* counting less than 10,000 inhabitants are exhaustively censused once within any five-year interval.**
The 35,750 *communes* counting less than 10,000 inhabitants, and which are home to half the population of France, are split off into five groups. Every year, the census survey procedure focuses on all the *communes* in one of these groups, and is conducted exhaustively across each individual *commune*. After the five years, the survey team recanvasses all the *communes* in the first rotation group, and so on.
The five rotation groups are balanced, i.e. *equally distributed* across ten or so demographics-based (population, gender and age-bracket) or housing-based criteria (number of housing units, number of primary residences). Samples are balanced at national scale and for each of the 26 French regions. This strategy, at equal-sized samples, yields greater accuracy.

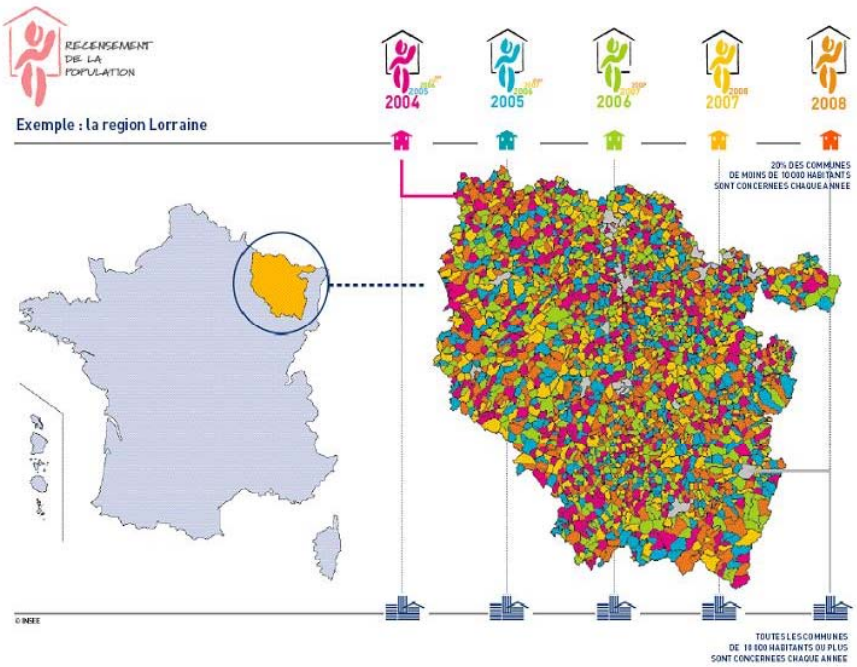


Fig. 1. The five groups of rotating communes for the Lorraine region

The 900 *communes* counting 10,000-plus inhabitants, and which are home to the other half of the French population, **run a census survey every year, but only canvassing a fraction of their populations.** Every year, the census survey canvasses 8% of housing units distributed across the territory of the *commune*. Hence, at the end of a five-year interval, 40% of the population of each

individual *commune* will have been surveyed, which is a sufficiently high rate to deliver robust data on the *commune* and its districts.

The sample frame for each large *commune* is built on the *RIL* buildings inventory. This inventory is an exhaustive list of all buildings (whether residential, administrative, industrial or business buildings) identified and mapped to an address. It is produced via a geographic information system capable of outputting a graphic representation of the data. The *RIL* was created based on the data from the 1999 national census program, and has since been updated regularly via administrative source files (building permits, local tax planning data) and postal files (Post Office address lists). Every year, the *RIL* is resubmitted to the *communes* for an assessment – each *commune* assessing their own territory – which is ultimately validated by the INSEE.

This commune-led reassessment is a pivotal process, as the *communes* possess intimate knowledge of their territories that can hone the INSEE's own data. Each *commune* is asked to provide input on *RIL*-updating workflows, and then on building stock. For workflows input, the INSEE sends the *communes* lists of addresses that need to be confirmed, i.e. addresses the INSEE wants to see either added to the *RIL* or deleted or corrected, so that the *commune* can issue a decision. These lists are sent out to *communes* twice-yearly, except when workflows are slow, in which case the INSEE may liaise with the *commune* to settle on a single dispatch grouping all the addresses requiring confirmation. The full *RIL*, i.e. the full stock of addresses, is reviewed in May-June time.
Quality metrics testing clearly indicates that *RIL* quality is better assured when the *communes* are actively enrolled in the review.
This expert review from the *commune* enables the INSEE to use the most up-to-date sample frame possible to extract the sample of addresses to be surveyed in the *commune* over the next January-February period.

The list of housing units surveyed for the census in each *commune* is sampled randomly from the housing buildings (addresses) list, according to the following sampling strategy:

- to avoid any 'clustering' effects *(see inset)*, the survey list features a subset of 'core addresses' that will be exhaustively censused over the five-year rotation, and that are consequently split over five yearly-censused groups. These core addresses are the buildings that house the higher number of housing units in each commune, and which meet the following criteria: buildings housing at least 60 housing units, and the full subset of core addresses in any one *commune* must not represent over 10% of the total number of housing units in the commune;
- new addresses are also comprehensively censused, as there are no datasets allowing them to be sampled; they are also split across five groups;
- the remaining 'other addresses' are split across five balanced groups, in the same way as the groups of *communes* that count less than 10,000 inhabitants, based on demographic or housing-stock criteria, with each group evenly distributed across the *commune* territory (so that any given street will include addresses from different groups); every year, the sample set of addresses targeted for census is sampled from the rotation group for that year, in such a way that the total set of addresses to be surveyed (core addresses + new addresses + 'other addresses') accounts for around 8% of the total number of housing units in the *commune*.

About clustering effects.
In any building or buildings registered to a given address, there is strong probability of between-housing unit and between-inhabitant similarity on variables concerning accommodation standards, social classification of the inhabitants, nationality, etc. Any statistical survey running estimates of these variables would almost certainly be influenced by whether or not this address was drawn into the sample. In demographics, this effect is known as clustering. The clustering effect gets stronger as the address matches to a higher number of housing units and to greater weighting in the geographic area from which the estimates are produced: the impact would be tangible at *commune* level and felt even stronger at the level of an 'IRIS' statistical block or any other zone encompassing this address. The fact that, over a five-year period, all the addresses above a certain size will get surveyed logically eliminates this drawback. The clustering effect does persist when dealing with smaller-sized addresses which are more homogeneous, but the impact is felt less.

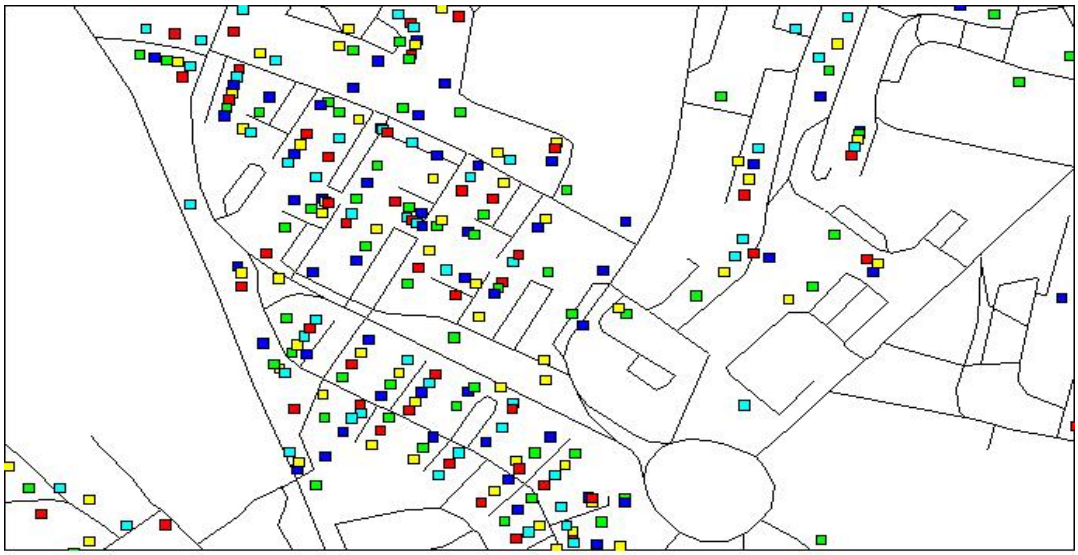


Fig. 2: the five address-segmented groups

The net result is that every year, we exhaustively survey a fifth of the *communes* that count under 10,000 inhabitants and 8% of the population living in *communes* of 10,000-plus inhabitants. This represents around 4.5 million housing units and 9 million individuals surveyed. **At the end of the five-year interval, 100% of the small-set *communes* and 40% of the population of each large *commune* will have been censused.**

In total, 70% of the French population is censused over a five-year rolling rotation.

Inset: ensuring balance between rotation groups

In *communes* that count less than 10,000 inhabitants, rotation group composition only changes if *communes* merge or split, or if the data collection system changes (switching from exhaustive coverage to data sampling, or *vice versa*). The large size of the groups practically eliminates any chance of balance shifts over time. Comparative analysis between 1990 and 1999 data has demonstrated that the balances remain highly stable between rotation groups. This exercise is scheduled to be repeated on the new census data. It should be underlined that if a loss of balance did occur, it would not affect commune-level population figures but rather the overall quality of an annual survey and the accuracy of the estimates.

In *communes* that count 10,000-plus inhabitants, rotation group composition varies as buildings appear or disappear. News addresses are split across the rotation groups in a way that maintains infra-group balance. Here again, cross-comparing 1990 data against 1999 data highlighted that the balances were very stable. This exercise is also scheduled to be repeated on the new census data.

All local communities (retirement homes, residential schools, penal institutions, religious community groups, accommodation centres, etc.) are comprehensively censused. Here again, data collection works to a rotation system spanning five years. Communities based in a *commune* that numbers under 10,000 inhabitants are canvassed during the year scheduled for the census survey on their *commune*. Communities based in a *commune* that counts 10,000-plus inhabitants are split into five groups and surveyed at five-year intervals along with all the communities in their group. The INSEE maintains a communities inventory for the purposes of organizing data collection and updating population records. Like the *RIL*, the communities inventory was created alongside the 1999 national census and is updated using administrative sources (registries on healthcare institutions, residential schools, penal institutions, etc.). This inventory is submitted to the *communes* for expert assessment.

Homeless people, people living in land-borne mobile homes, and boat-dwellers ate comprehensively censused once every five years.

Data collection remains organized around a system of hand-out-hand-back questionnaires.
Although the sampling system represents a major innovation, data collection is performed according to the same method as for national census programs: hand-out-hand-back by census enumerators. Each household is given two separate forms to fill out: a housing schedule and personal schedule. These questionnaires are very similar the one used in previous censuses.
The first questionnaire, called the *housing schedule*, records the list of inhabitants at the household, through fifteen or so questions polling items relating to the standards and characteristics of the housing unit and the household's vehicle ownership data.
The second, called the *personal schedule*, features 25 questions canvassing age, gender, place of birth, nationality, place of residence five years earlier, education attained, occupation, college or workplace.

Unsurveyed households listings to account for non-response

Unsurveyed household listings (*FLNE*) represent a major innovation helping the new census account for non-response. The system works to the following principle: for primary residences that could not be surveyed (residents impossible to locate, away from home for long periods, or who refuse to respond), the enumerator completes an unsurveyed household listing (*FLNE*), stating the number of residents in the household so that the population figures can be rectified. **The population of these households is therefore registered in the census counts.**

This means that either a housing schedule plus one or more personal schedules or an *FLNE* get collected from every primary residence in the *commune*. This input is checked at collection and then double-checked at the INSEE.

3. Procedure for calculating official population figures and statistical output.

The INSEE is required to meet two conditions for publishing population figures:

- publish population figures **every year for every *commune* (statutory obligation)**
- for the purposes of achieving equality in processing population figures for different *communes* and quality in surveying populations spanning different *communes* (coherency between figures to be totalled), **this population figure has to concern the same year for everyone**. It would be wholly unreasonable to issue the 2004 figure for one *commune* and the 2008 figure for another.

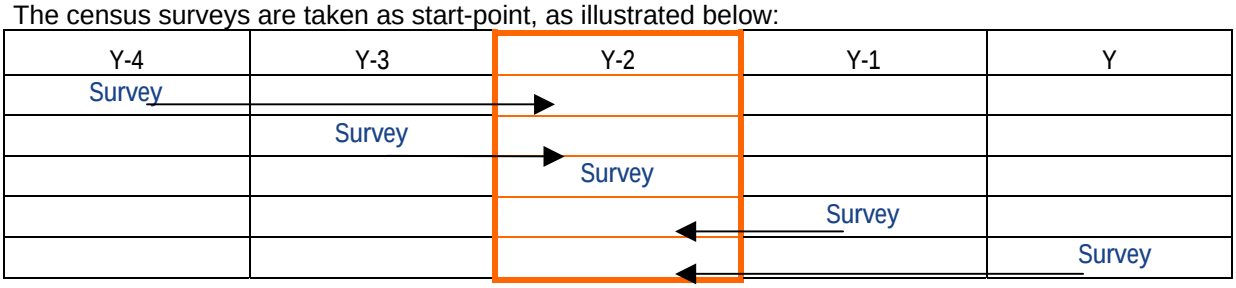
The INSEE calculates these populations as set out below:

The householder population

For *communes* counting 10,000-plus inhabitants, the householder population figure is based on a rolling average calculated from samples taken in five successive years. These five samples from years Y-4 to Y are aggregated to calculate an average population per household representing the situation at the midpoint of the period, i.e. year Y-2. This ratio is then multiplied by the number of housing units at the start of year Y-2, as recorded in the *RIL* buildings inventory, to give the householder population of the *commune*.

In order to also integrate the *communes* counting less than 10,000 inhabitants, the population count for the median year in the 5-year rolling rotation has to be corrected to fit with the figures for the *communes* counting 10,000-plus inhabitants.

The census surveys are taken as start-point, as illustrated below:



For *communes* censused in Y-2, the result used is the count from the census survey.
For *communes* censused in Y and Y-1, the Y-2 population is estimated by interpolation between the results of the last two census surveys, i.e. between the last census survey and figures for the 1999 census when the system was first deployed.
For *communes* censused in Y-4 and Y-3, the calculation is based on an extrapolation between the result for the census survey and Y-2. This extrapolation step draws on data from housing tax registers, which give a picture of the patterns of change in the number of housing units per *commune*. The output data is further processed so as to factor in the incremental shift between the growth in the housing units count and growth in the residents count. This incremental shift recorded between the last two censuses is applied to the changes measured by the housing tax register to give the change in the residents count.

The non-householder population

The community population figure is corrected to January Y-2, i.e. the same date as the householder population figure. If the communities in the *commune* were censused in Y-2, then it is the figure given by the headcount that is used. If the community population were censused in Y-4 or Y-3, then the figure is updated by adding the population of new communities and striking off the population of communities lost to follow-up. A new community is qualified as a community that has appeared since the latest records collection program; these new communities are pinpointed through administrative source files (registries on social/healthcare institutions, residential schools, penal institutions, etc.). The changes to be integrated are determined based on the communities inventory. If the communities in the *commune* were censused in Y-1 or Y, then the population figure for these communities at January 1st Y-2 is calculated by interpolation between the figures for the last two community censuses.

Population figures on homeless people, people living in mobile homes, or people whose main address is given as a hotel are not updated, but instead re-posted without change for each of the four years following the collection year. New collection data replaces the previous data as soon as it becomes available.

4. Rationale underpinning the reform

There are two factors driving the census reform: a need for more regular, more up-to-date data, and a move to more efficiently pace the human and financial burden of a national census program.

End-users of INSEE data output have been increasingly voicing the need for more up-to-date data. National census programs conducted at increasingly long intervals (nine years between the last two, which were led in 1990 and 1999) were losing the ability to track demographic trends in France: social or familywide relocations, urban sprawl, city planning, etc. One of the main groups pressing for reform has been local authorities, which are the leading users of census data, especially since progressively inheriting a greater share of decision-making authority under national decentralization policy: greater governance responsibility has been extended to *communes* and *commune* clusters on public transport issues; to *départements* on social welfare management, assistance services to senior citizens, and middle schools investment and operational management support; to the regions on issues like vocational training, spatial planning, high schools investment and operational management.

To properly exercise their jurisdictional responsibilities, local authorities need regular, up-to-date demographic statistics in order to build decisions and assess their impact. The new census scheme meets these needs by providing detailed statistics output at every geographic scale, from country-wide down to *commune* and district level.

In 1996, the Economic and Social Council report on 'Region-specific demographic features and spatial planning' was critical of the growing lag between French general censuses.

The increasingly dispersed general censuses also suffered from the equally problematic burden of heavy workload and burgeoning financial costs. Budgetary constraints had already forced the census initially scheduled for 1997 to be postponed to 1999. By pacing the workload more efficiently, the new system makes it possible to spread the census costs and work to a sustainable, annually-rolling budget. Enumerators and census staff (both at the *communes* and the INSEE) can better consolidate their professional competencies, as they are called upon more often. The new system ushers in smaller-scale, more manageable census campaigns that make for more efficient organization and improved process control. Furthermore, working to a sample frame (the *RIL*) improves monitoring and follow-up on the data collected (see the section on data pooling below).

B/ Compliance on international quality criteria

5. Compliance on UN-defined essential census features

A UN advisory on census-taking in the decade spanning 2010 set out a series of principles and guidelines on what a population census should contain ('core variables'), defining five 'essential features': individual enumeration, universality, ability to output data on small-scale unit areas, simultaneity, defined periodicity. The new French census system meets these criteria.

'Individual enumeration'
Information is collected separately for each person or household surveyed. A detail-file (featuring one record per person or per household surveyed) will be issued every year from 2008 onwards, collating the file data from over a five-year survey period so that the data can be cross-tabulated, in the same way as for mass national census programs. Each of the 9 million people and 4.5 million housing units surveyed each year are included in the file, complete with the sample weighting. The detail-file therefore integrates each individual person.

'Universality'
Every *commune* that counts under 10,000 inhabitants is surveyed every five years. The entire population of each *commune* is censused. This guarantees the same level of universal coverage as for mass national census programs.
For *communes* that count 10,000-plus inhabitants, the entire territory of the *commune* is surveyed within a five-year period, as stipulated by law. The sample frame covers all the residential buildings within the commune, thereby fully meeting this constraint according to the principle of universality. The sample frame is re-updated annually by cross-checking against Post Office address lists, building permits and/or local tax planning data. This updating process is led in tandem by the INSEE and the *communes* themselves.
The data collected in five-year blocks therefore covers the entire population and the full territorial unit area covered by each *commune* of 10,000-plus inhabitants.

The five-year cumulative sampling coverage in large *communes* reaches up to 40%, which guarantees the accuracy of the results recorded. This is an improvement on the national general census program, where certain output variables were based on a sample of people or housing units only. There are also countries that only mail out the 'long-form questionnaire' to a 15 or 20% fraction of the population.

The net result is greater assurance on census exhaustivity provided through the unsurveyed household listings (*FLNE*), and the better-quality data collection.
The survey sampling system and year-by-year census fieldwork makes for better quality data collection as the system harnesses more focused efforts from both INSEE agents and *commune* staff. Furthermore, in the *communes* with 10,000-plus inhabitants, census enumerators work to a list of addresses to be censused, making their work more efficient (see page 8). Populations living in collective households are exhaustively surveyed, as homeless people, people living in land-borne

mobile homes, and boat-dwellers. To round off, deploying the *FLNE* makes it possible to estimate the number of nonrespondents. The net result is that data is collected on every single person living in the country's borders.

'Ability to output data on small-scale unit areas or small-population groups'
Compliance on this criterion is a logical extension of compliance on the principles of universality and individual enumeration. The sampling coverage rate, at 100% in *communes* that count under 10,000 inhabitants and 40% in *communes* of 10,000-plus inhabitants, coupled with the data organization system based on weighted personal data files makes it possible to output detailed statistics at highly-granular geographic scale or covering small-population groups. The five-year datapooled file covers around 45 million people and 22 million housing units, allowing for exceptionally granular tabulations.

The census will output detailed annual data on each *commune*, and down to IRIS-level statistical blocks (districts housing around 2,000 inhabitants) in more major *communes*.

Either way, for small-sized geographic blocks, it was decided that any yearly statistical output, even if it lacks accuracy due to survey-related uncertainty, would prove a better option than having to resort to often wildly outdated information. For instance, the rolling census system would prove far more beneficial for tracking urban redevelopment projects than a general ten-year census scheme. This was a topic debated during the census consensus-building phase, as certain end-user groups leaned towards a more exhaustive census program enabling more granular analysis than the theoretically less accurate annual data collection scheme.

'Simultaneity'
Census surveys take place on the same data every year (and starting on the third Thursday in January) in all the *communes* earmarked for that year.

The data collected each year is corrected to a single date, which is set at the midpoint of the rolling five-year rolling rotation.
In the larger *communes*, the five samples are totalled together, before 'fitting' the individual weightings to the number of housing units sampled from the *RIL* buildings inventory for the median year.
In less-populated *communes*, the procedure is to interpolate or extrapolate between the census survey and the reference date for the population. The extrapolations are consolidated by drawing on data from housing tax registers to gain insight into the patterns of change in the number of housing units per *commune*.
The survey period selected (five years) dictates that the interpolation/extrapolation radius never extends beyond two years, which minimizes the potential for drift.

'Defined periodicity'
The fifth essential feature is a defined periodicity: censuses need to be held regularly, every N years (every five or ten years for instance). The French census scheme outputs results every year, and therefore comfortably meets this defined periodicity requirement.
It also offers end-users annual data, enabling them to bypass the date constraints imposed by the classic quinquennial or decennial census window.

6. Compliance on the quality criteria shortlisted by the EuroStat code of best practices

EuroStat has established a list of quality criteria that have been borrowed (with a handful of terminology edits) for the European statistics code of practice adopted by the Statistical programme committee on 24 February 2005. The new census scheme is fully compliant with these requirements, and the methodology has been recognized in 2008-published EU census legislation alongside general census schemes or population and housing registers.

'Relevance':
The reform is designed to output statistics data that better meets end-user expectations, namely through regularly updated data. A panel of consensus-building initiatives, led during the census groundwork phase and in some cases continued progressively throughout operational rollout, makes it

possible to factor user needs into the design and development of statistics products and publication services (see parts 1 and 3).

'Accuracy:
Sampling-related random error (maximum acceptable 0.05% for Metropolitan France) shall be below the data collection error commonly recorded for general censuses (often around the three–point mark in most of the countries that run general census programs) and also below the level of error induced by data ageing (the mean average annual population growth is around 0.7% per year).

Accuracy of population estimates for major *communes*

Population	Coefficient of variation (%)
10,000	1.1
15,000	0.8
20,000	0.7
30,000	0.5
40,000	0.4
50,000	0.4
100,000	0.2
150,000	0.2
200,000	0.1
300,000	0.1
500,000	0.1
1,000,000	0.1

Accuracy is essentially dependent on the size of the sample population. This table gives the coefficient of variation for the population count according to population size, i.e. the standard deviation relative to the population figure. The coefficient of variation is a measure of the relative accuracy of the survey.

Considering data collection error, the new census scheme offers better process control (see the section covering universality), even though 'zero error' remains out of reach. Furthermore, the data adjustment options offered by the *FLNE* listings together with extrapolations based on the *RIL* offer better control of undercount due to omissions. That said, although the scheme provides improved coverage, it also implies a concomitant dip in overcount performance. Overcount is actually tougher to control in rolling census systems based on sample surveys than in an exhaustive, one-time mass census, where duplicates are easier to identify.

A critical point on data accuracy concerns the measurable level of **accuracy of time-course patterns**: just because the census is made available every year does not make it a source flagging cyclical trends. Short-term variations have to be interpreted with caution, and under certain conditions:

- Shifts in the population figures or housing stock can usually be analyzed through year-to-year datasets;

- Analyses on patterns of change involving statistical variables shall be run by comparing two model years spaced at least five years apart so as to be certain that the samples are totally unconnected. This cuts the scale of the comparative intervals down to half the values for general census programs, but does offer greater flexibility in the choice of useable sample years.

'Timeliness and punctuality':
This is one of the strong points of the new census system.
Timelines of the data output release dates is an improvement over general census programs (and this was one of the objectives targeted by the reform): the new census generates statistical data stretching

back two to three years at most, whereas data from general census schemes is between 18 months and ten years old.
Punctuality is written into law (the first datasets delivered before end-2008, with annual periodicity thereafter); compliance with the data publication schedule is a pivotal factor for success, given that the system operates to an annual release schedule; *commune*-by-*commune* population estimates that were delivered at the end of the year factored in the data collection at the start of the year for all the *communes* counting less than 10,000 inhabitants surveyed in 2004, 2005, 2006 or 2007, plus data from around 850 *communes* counting 10,000-plus inhabitants for which the INSEE considered that the estimates obtained after four data collection campaigns was sufficiently robust to merit publication.

'Accessibility and clarity':
The statistics are mainly disseminated through online sources, which ensures very satisfactory accessibility. The *communes* are given access to their data before public release, at which point they also have the option of releasing the data to their constituency. Finally, the press gives broad coverage of the main census results.
The new data collection and processing methodology makes data format and user guidance critical challenges. The INSEE is therefore mobilizing a significant raft of resources to cover these needs as part of an ongoing drive set to span several years, given the vast range of uses for census output.

'Comparability':
Comparability is a component with several overlapping dimensions:

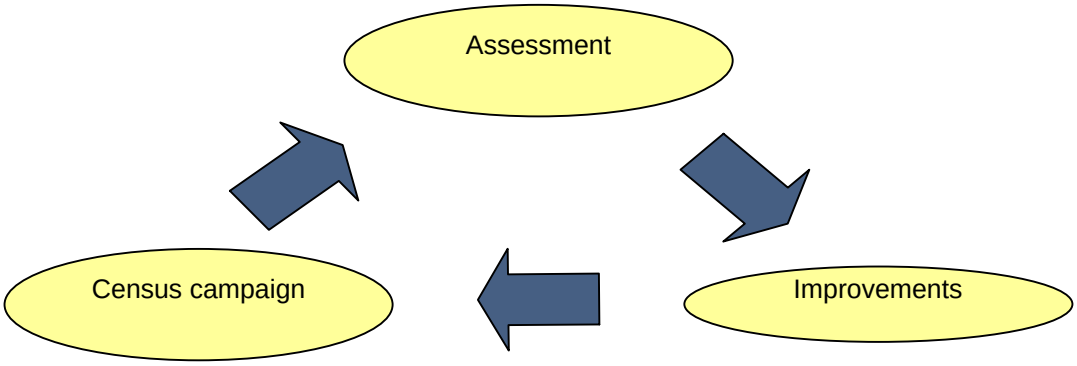
- Comparability over time: the questions asked of respondents have seen relatively little change compared to previous censuses, and these changes have been documented; furthermore, the fact that the new method has been designed based on annual sample surveys eliminates one of the main comparability problems hampering general one-time census, namely the effect of changes to data collection protocols. From now on, with regularly-paced surveys employing the same protocols, the new census is better placed to guarantee statistical comparability.
- Comparability between countries: this is now guaranteed at UN and EuroStat level through compliance with the 'core variables' policy: the internationally recommended core questions all feature in the French census and have been compiled based on internationally standardized concepts.
- Comparability between cross-national geographical regions: infra-national comparability will be guaranteed by annual statistical outputs all indexed to the same dates. However, end-users will need to take certain precautions when interpreting business cycle variables such as unemployment figures depending on the area breakdowns considered. The French National Statistics Council [*Conseil national de l'information statistique*] task-force report on published statistics gives a number of illustrations highlighting this issue, and proposes methods for getting the best out of the statistics available.

'Coherence':
Statistical coherence is ensured by a series of adjustments designed, among other things, to correct the small number of inconsistencies between variables that will inevitably creep in.
Turning the focus to coherence with other local data sources, one of the key challenges is going to be to 'pool' survey data with other local sources: the time when the population census was the only source of localized data is now well and truly relegated to the past.

C/ Census quality management policy

7. The global approach

Running annual census campaigns makes it possible to integrate quality assessments, as illustrated in the diagram below:



Process improvements earmarked for a census campaign N are not necessarily implemented at N+1 and onwards, but instead at N+2 or N+3. This is to free up the time needed to investigate the different questions, lay the groundwork to the changes and usher them into practice at every step: protocols, software applications, notifications, training, etc. Some of the most straightforward-looking changes can actually generate heavy design costs (to safeguard process-wide coherence) and/or implementation costs (editing software applications, printouts, and organizational protocols, and training the agents concerned).

Process improvement is piloted through an approach centred on systematic assessment of the full set of processes:

- at the INSEE, through assessment reviews canvassing for feedback from all the actors involved, including *commune* coordinators. These qualitative and quantitative assessment reviews are focused on process relevance, choices on the use of applicative functionality, the organizational measures implemented, and workload burdens. The reviews are conducted via questionnaires but are often supplemented with one-to-one meetings;
- in cooperation with subcontracted services tasked with data input, who produce a full operations report on campaign completion;
- and via the National census program review committee, which brings together representatives from the INSEE, the *communes* and user groups (see section 3) in meetings held two or three times a year. This commission was created within the French National Statistics Council (CNIS), and meetings are chaired by a senator. Alongside the INSEE and the *communes*, the census user groups include administrations, councillor associations, researchers, and so on. The Commission assesses census administration and data collection processes and tables revisions to protocols, in addition to voting on suggested changes to the regulatory texts governing how census campaigns are to be organized;
- finally, process quality metrics and quality control checks are carried out at every stage in the program (spanning sample frame coverage, data collection, keying and processing statistics, publishing the data).

How quality monitoring changes in relation to a general national census

The sample survey scheme paired with an annual data collection program drastically cuts the volumes to be processed (seven-fold fewer questionnaires than under a conventional census exercise) and the workforce mobilized (five-fold fewer *communes*, six-fold fewer enumerators). Consequently, the INSEE, the *commune* coordinators (appointed by *commune* mayors to run local census action) and the enumerators are able to better focus their efforts, resulting in improved data collection quality and sharper checks and controls.
In *communes* with 10,000-plus inhabitants, each enumerator works to their own list of the housing units they have to cover. This is a crucial factor, as the fact that the enumerator knows exactly where they have to go makes their work more efficient than the 'square search' technique employed in general censuses.

8. Data control and validation, process by process

Each individual census process features a census quality control component. The assessments outlined in the previous paragraph are further supplemented by quality control operations run during the key phases: quality control on the sampling frame; quality control on data collection; quality control on data keying and coding; quality control on data earmarked for publication.

The INSEE therefore implements inventory control through field surveys and implements field data collection control by cross-analysis against administrative source files, which are reconciled, if needs be, with a field survey. Furthermore, data keying and coding operations are controlled through double data entry and double coding on a schedules sample, with both operations being followed up by variance analysis.

8.1. The population figures

Commune population figures are calculated based on the following inputs:
- housing stock compiled from the *RIL* buildings inventory;
- the files compiled through the various data collection operations;
- the communities inventory;
- the housing tax register.

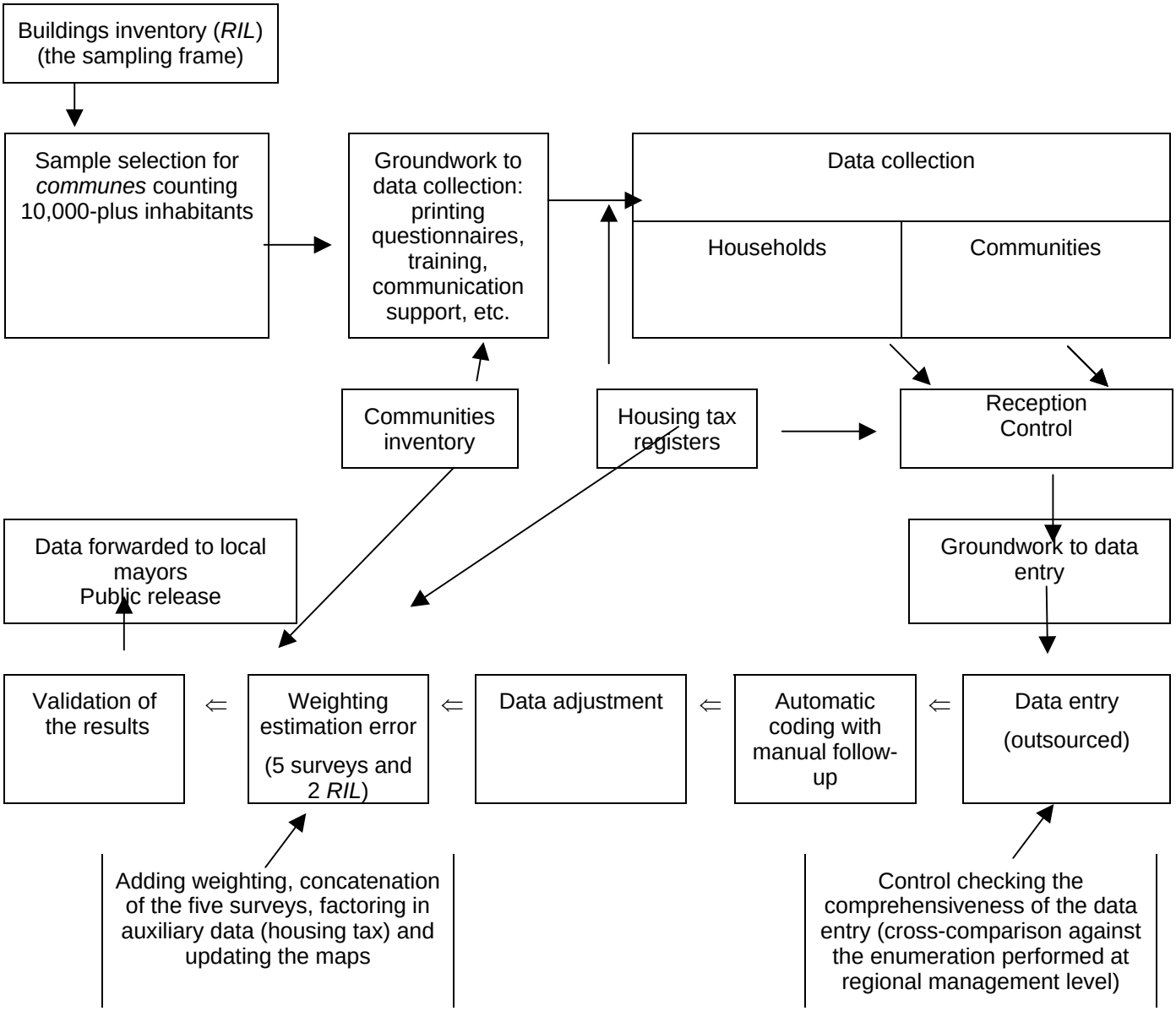
Each of these inputs goes through the validation process, which is led end-to-end from production through to processing, as it is easier to correct errors when they are identified early.

The validation process also includes inquiries led on two factors that play a role in data quality and interpretation: patterns of change due to conceptual differences in census designs, and the effect of data adjustments. In *communes* that count under 10,000 inhabitants, given that sample survey results are put through weighting and extrapolation steps, the validation process also focuses on the effect of 'core addresses' capable of introducing clustering.

Data validations has two main objectives: one is to provide guarantees of the statistical quality of data output intended for public release, while the other is to understand the patterns of change identified, estimate their robustness and, accordingly, issue precautions for use. This second factor becomes particularly important when a given *commune* is strongly affected by differences in conceptual design.

The schematic diagram below illustrates the sequencing of the various steps in the census process.

SEQUENCING THE CENSUS STEPS



a) The housing stock compiled from the *RIL* buildings inventory

Census quality hinges on the quality of the *RIL*: a sampling frame that lacked too much data would not offer the required level of representativity, whereas a sampling frame that carried too much surplus data would just bloat the data collection (by generating waste). To conclude, it takes a good balance between shortfall and surplus to calculate bias-free estimates when processing the data.

National and possibly even regional-level quality metrics are run on the buildings inventory (*RIL*) via an annual field follow-up survey led by INSEE investigators.

The aim is to measure *RIL* completeness in terms of deficit or surplus addresses in the *residential* category and the associated housing units. The metrics are designed to provide guarantees on matchup between all the addresses inventoried in the *RIL* and the addresses found in the field.

Procedural implementation of the metrics quality follow-up survey

The survey is focused on a selected block sample. A block sample generally spans a block of houses, but can also cover a section of a large residential complex. Each block is surveyed by enumerators who conduct an on-site follow-up on the addresses listed in the *RIL* as being located in the sample block: residential housing units as well as other address categories (institutions, facilities, or 'other'). They double-check that all the addresses listed match with the reality in the field, with priority going to residential housing. Any listed residential address that is not found in the field is registered as a surplus, as is any address that the field investigation reveals to be only a secondary entrance. Any residential address that is found in the field but not listed is registered as a deficit. The survey team also runs checks on the address system, the number of housing units, fuzzy overlap between categories, and addresses with several entrances.

Again, the primary objective of the survey is to run controls on the completeness of the residential *RIL*, in terms of addresses and housing units, and most importantly to detect address deficits.

The 2007 block sampling campaign covered 340,000 addresses and 900,000 housing units.

Results of the metrics quality follow-up survey

Shortfall and surplus residential addresses and housing units listed in the RIL

		RIL 2003	*RIL 2004*	*RIL 2005*	*RIL 2006*	*RIL 2007*
Addresses	*Deficit*	*2.3%*	*1.9%*	*1.3%*	*1.3%*	*1.4%*
	Surplus	*2.0%*	*1.9%*	*1.8%*	*1.7%*	*1.4%*
	Surplus-deficit balance	*-0.3%*	*0%*	*0.5%*	*0.4%*	*0%*
Housing units	*Deficit*	*2.3%*	*1.9%*	*1.1%*	*0.9%*	*1%*
	Surplus	*2.1%*	*2.3%*	*1.8%*	*1.4%*	*1%*
	Surplus-deficit balance	*-0.2%*	*0.4%*	*0.7%*	*0.5%*	*0%*

Source: INSEE surveys on RIL quality: 2003-2007

b) Files sampled through the data collection process

The quality of these source files hinges on two key factors: the quality of data collection and the quality of data entry. Indeed, the data adjustments made are designed to improve quality by correcting nonresponse and uninterpretable response. The coding step is also targeted by a quality control measure.

The data collection itself often proves the potentially weakest link in the statistics process. Data collection quality hinges on sharp, meticulous procedures, but equally important as factors are positive endorsement from the respondents (which governs the sincerity of the responses given) and co-option from the people collecting the information, i.e. the enumerators. Meticulous groundwork (pre-census field surveys, training for agents) and quality control measures provide the necessary quality

guarantees. Both the INSEE and the *commune* correspondents run controls throughout the data collection phase. Once the field collection has been completed, the INSEE takes over process control to double-check and correct the data collected.

Data collection groundwork and in-process controls

Data collection is managed by the *commune*. Procedural rollout follows a clearly-defined protocol, implemented by the enumerators and supervised by the *commune* head coordinator and the INSEE census supervisor. The enumerator handbook sets out the assignment tasks. *Commune* coordinators and census supervisors are given detailed documentation. The *commune* runs in-process and post-process data collection checks to make sure the enumerator canvasses every single housing unit scheduled for census. Not a single address or housing unit can be allowed to slip through the net.

The supervisors also carry out controls during their time in the *communes*, checking that the procedures are being followed (confidentiality, working to deadline) and that the questionnaires already checked-in have been satisfactorily completed. This enables them to screen for the most common types of errors. At completion of the data collection phase, when the questionnaires are handed over to the INSEE, the supervisors give a spot appraisal of the level of data collection quality in the commune, indicating any difficulties that may have been encountered.

In addition to this core quality assurance system, the *communes* counting 10,000-plus inhabitants also deploy a more elaborate census quality follow-up and control system.

Upstream of the collection phase, the addresses identified for the sample to be surveyed are quality-controlled through a two-stage procedure: the addresses and the number of housing units are validated by the commune's central census team before the enumerator goes on to field-check the addresses and count up the housing units (this phase also gives the *commune* an opportunity to test the performance of the enumerator). If the enumerator and central team do not reach consensus, a second sequence of field-checks is run (the *commune* team, depending on the resources at its disposal, may opt to allow a variance threshold). The outcome is that once the data collection phase is actually launched, both INSEE, enumerator and *commune* coordinator are all agreed on the address identified, on its status as inhabited, and on the number of housing units to be censused. The outcome of this methodical upstream control phase is that the enumerator has a more 'comfortable' data collection assignment to execute.

In all *communes*, the ongoing follow-up on data collection operations and enumerator performance makes it possible to track census form hand-out and hand-back progress, identify any difficulties encountered by the enumerators, and highlight any under-provision. The INSEE provides support on this follow-up process by issuing the *communes* with a collection tracking and control tool, but *communes* may opt to develop or outsource the development of a proprietary tool if they are looking to work with more advanced functions. The first documents that the enumerator checks in are scrutinised by analyzing the quality of how the forms are filled in order to assess how well the enumerator has understood their task instructions and their level of commitment to the collection assignment.

Some *communes* hold a roughly two-hour weekly review with each enumerator. This policy is designed to double-check that all the addresses, all the housing units listed for each address and all the residents housed there have all been censused. The enumerator works to weekly performance targets (numbers of forms handed out and documents checked in). As this management phase draws to a close, the final remaining collection issues are analyzed. There are several key performance indicators: number and percentage of *FLNE* listings (to be established as late on as possible, in the operations wrap-up phase), housing vacancy rate, quality of the census records logbook.

After collection: Controls led by the INSEE

Once the INSEE has received the documents forwarded on by the *communes*, it conducts desk reviews followed by field reviews. The control protocol is geared specifically to the new-format census survey system. The lower volumes of annual collection data to be processed and the lower number of *communes* censused make it possible to conduct more extensive checks than are run under national census programs. The housing tax register plays a key role in the desktop reviews, making it possible to double-check the completeness of the collection data for a given *commune* or collection area. Both the desktop control checks and the field control checks led are led 'hands on', and the errors detected are all corrected.

They are selective, as they stem from a series of successive screens sifting through the data:

- a first-in control when the questionnaires are received are checked in yields an appraisal of data collection performance in each commune;
- for around 15% of the total *communes*, i.e. 30% of the national population, the outcome of this appraisal is an extensive desktop control;
- any cases that remain unsolved following this desktop review are then targeted for field controls led by the INSEE (in 2007, around 28,000 field controls were led, spanning 8% of French *communes*).

Desktop-led controls

The first step, termed 'Questionnaire reception and check-in', focuses on the documents sent in from all the *communes* involved in the census survey for the current year. INSEE clerical staff use a barcode system to count in the questionnaires. They double-check addresses, record uncompleted questionnaires, and so on, in what is known as the 'code-scanning' phase. This major processual innovation pioneered through the new census system enables every single schedule to be identified and tracked throughout the processing chain. The process also generates the input data for per-*commune* and per-assignment area indicators on changing trends in headcount and housing numbers, structural changes in housing stock (primary residences, second homes, seasonal housing, unoccupied housing units), and the proportion of unsurveyed household listings (*FLNE*). The step concludes in an expert review on each *commune* that factors in these quantitative indicators alongside qualitative indicators on data collection performance filed by the supervisors when they give their appraisal on the quality of the documents checked in. The indicators are collated via a scoring system to produce a census 'quality' score on each commune, and to trigger advanced quality control review on the *communes* posting the worst scores (covering roughly 15% of all *communes*, i.e. 30% of the total population sample).

The second step is advanced desktop review.

This advanced review may focus on an entire *commune* or just a section. Centred mainly on tax register source files, it will be purposed, depending on the issue highlighted by the indicators, towards:

- the completeness of the housing unit count:
  - whether the address has been mapped;
  - cross-checking the housing unit counts;
  - cross-checking the occupants names (*communes* with 10,000-plus inhabitants);
- the percentage of unsurveyed household listings (*FLNE*):
  - checking the *FLNE* based on the 'housing category' data in the housing tax register,
  - projected headcount resident at the housing unit;
- structural checks per housing category (primary residence, second home, seasonal housing, unoccupied housing unit):
  - cross-comparison with the data recorded in the housing tax register.

The advance desktop review may involve running phone interviews with residents.

Taken together, this raft of quality controls results in either the data collection being validated or errors (such as second homes being wrongly recorded in the census as primary residences) being highlighted. Any errors detected are then corrected.

Step three: field follow-up controls

In the event that the desktop review is unable to deal with a number of discrepancies, the decision will be taken to run in-field follow-up controls on the *communes* concerned. In this scenario, the *commune* is informed of this decision, but does not have access to listing of the addresses or housing units concerned. These in-field follow-up controls are run on *communes* considered as underperforming on census quality (either across the entire commune, which is in fact a fairly rare occurrence, or on a section of the commune) following the INSEE-led desktop review. In the vast majority of cases, the field follow-up controls confirm the data input collected, but there are occasions where certain errors

are detected and thus corrected, in which case the questionnaires are edited and the *commune* is notified of the corrections made. These quality controls are equally valuable for adapting any adjustments made and editing the data collection or data-checking protocols for the following year.

Field follow-up controls are led in April and June by specially-trained INSEE enumerators.

In 2007, over 28,000 field-follow-up controls were led in a total of 691 *communes*, i.e. covering 8.5% of the *communes* involved in the census survey that year. Filed follow-up controls were led in close-on a third (31.7%) of all the *communes* counting 10,000-plus inhabitants. As the priority for these controls is to focus on the most problematic cases or on proven discrepancies, there is no rationale for including them as part of the statistical output for the full data collection process.

Field follow-up control operations focus on either addresses or housing units.

Controls on addresses are run on addresses that do not appear to match up to the addresses geocoded in the sample population (*communes* counting 10,000-plus inhabitants), or on addresses that may have been skipped or indeed that are not found to be recorded in any other source (*communes* counting under 10,000 inhabitants). The housing unit is double-checked at the same time.

6,400 address-matching controls were led under the 2007 field follow-up campaign, which equates to 1.3 per 1000 addresses in *communes* counting under 10,000 inhabitants and 7.3 per 1000 addresses in *communes* counting 10,000-plus inhabitants.

Controls are also run on housing units whenever the unit is filed in an *FLNE* or when there are questions over which is the right housing category (primary, second, vacant or seasonal) to apply. The headcount is double-checked at the same time.

Around 22,000 housing unit-matching controls were led under the 2007 field follow-up campaign, i.e. one in 200 housing units.

A large majority (79%) of these quality controls were focused on housing units filed in an unsurveyed household listing. It is through these controls that the statistical output (population figures and detailed breakdown) is able to account for the proportion of the population that the enumerators were unable to canvass directly. This is largely the case for *FLNE* established on primary residences (4 in 5 cases), and the headcount estimate on the number of people resident in these housing units proves robust (headcount estimates correct in 4 out of 5 cases, and there is an equal fraction of overcount and undercount in relation to the numbers recorded in the *FLNE*).

The other housing unit follow-up controls are designed to check housing category, which proves confirmed in around 75-80% of cases depending on the type of housing unit. Again, it should be underlined that these checks are primarily run on cases where inconsistencies are either detected or suspected.

The take-home message from this overview on the data collection controls carried out is that the quality control system deployed is one of the key advantages of the new method: the smaller-scale census survey campaigns combined with successive screens sifting through the data translate into far more efficient quality controls than under a conventional census program.

Validation of the data entry

Data entry is outsourced to an independent service provider. Data is entered via optical scanning: the questionnaires are first scanned and then run through pattern recognition software (digits, letters, checkboxes, barcodes). The process is rounded off by manual input of handwritten data the hardware was unable to recognize. A detailed set of specifications defines the level of quality performance the INSEE expects from the outsourcer. The specifications extend from data protection in transit and once delivered to the service provider (physical and computer system security, plus nondisclosure agreements from all agents involved) through to the data entry quality performance (number of documents processed per assignment area, maximum error rate per variable). Finally, the documents are counted up and the total is cross-checked against the first count carried out at the INSEE in the code-scanning phase (see above), looking for deviations.

The net result is that during the process, the records (personal schedules and housing schedules) are counted a total of three times: once by the *commune* to wrap up the data collection step, once by the INSEE during the document check-in and control phrase, and once again by the service provider. Cross-comparing the three counts guarantees an extremely high level of process quality on the number of schedules processed.

Finally, the data entry itself is also quality-controlled via a double data entry strategy (see inset)

Double data entry to check the quality of the variables entered

Data entry error rate is assessed via a second-pass data keying procedure on a schedule sample set performed by a second service provider, working from scanned images of questionnaires in the sample set while working to the same data entry ruleset. Any deviations between the two sets of data entered are analyzed at the INSEE, where error rates are calculated for the primary service provider's coding performance on each variable. Double data entry verifications are carried out throughout the data entry campaign – a measure that makes it possible to correct the data entry protocols for the following data batches.

A series of meetings and feedback sessions are scheduled with the service provider on a regular basis in order to prepare against or swiftly resolve problem issues. Following up on lessons learnt from the double data entry exercise, the provider is asked to make adjustments or improvements, either mid-campaign or in the inter-campaign interval if the improvements prove extensive. For instance, optimizations led on the pattern recognition software made it possible to significantly cut the data entry error rate on dates of birth. Indeed, it is important to underline that even on data as apparently straightforward as dates of birth, hoping for zero error rates would equate to wishful thinking: as the source data is handwritten, there will inevitably be interpretative mistakes.

Corrective adjustments are implemented as a measure to improve the quality of the files sampled through the data collection process. The measure involves analyzing total and partial nonresponse, reconverting it to give valid information, but it also entails correcting schedules that contain inconsistencies (such as a 95-year-old claiming to be actively employed). The procedure employed, known as 'hot-deck' imputation (previously used in past censuses), consists in substituting the missing data with a 'donor' presenting a similar shared profile. Here again, the corrective adjustments progressively iron out issues pinpointed in previous results, which equates to an efficient quality improvement system.

In particular, adjusting the *FLNE* listings makes it possible to produce an estimate for *commune* population calculations on the population of primary residences for which the enumerator was unable to canvass residents (residents impossible to locate, away from home for long periods, refusals). These are housing units for which the enumerator establishes an *FLNE* stating the resident headcount (see section page 4 on *FLNE*s). For primary residences for which the number of occupants has been entered (more than 8 in 10), the figure 'imputed' for the household is the number of occupants stated in the *FLNE*. Imputations are also entered in the other *FLNE*s following a method designed to reproduce the structure recorded as a national average for each household size in fully-completed *FLNE*s.

c) The communities inventory.

Like the *RIL*, this inventory was compiled from data collected under the 1999 census and updated via administrative source files (registries on social/healthcare institutions, residential schools, penal institutions or institutions tasked with the legal protection of minors, and so on). Like the *RIL*, it is forwarded on to the *communes* for expert review just before the census survey. The *communes* receive the listing of their communities and are asked to indicate any communities that have may have disappeared, escaped identification by the INSEE, or changed address. The INSEE then decides whether to make the additions or deletions based on the feedback from the *communes*.

When the collection campaign is launched, the communities enumerator is tasked with running a final validation on the communities listing.

Finally, the INSEE cross-compares the *RIL* and the communities inventory on an annual basis. The first strand of the process is to confidently geolocate each community in the inventory, with the *RIL* acting as reference listing for addresses and their geographic profile (x-y coordinates, and whether it

forms part of a district, a canton, or an infra-*commune* area). The second strand involves cross-comparing the two inventories, checking for duplicate entries or omissions (most importantly when a community mutates into a standard set of housing units or *vice versa*).

d) The housing tax register

Every year, the French Directorate of Public Finances (DGFiP) issues the INSEE with a file summarizing all of the premises registered to pay housing tax. This file is first screened to weed out premises that are not qualified as residential (primarily garages and car parks).

The file is then utilized for two purposes:

- data collection follow-up and process control, on the *communes* involved;
- serial listings tallying total housing units, the number of primary residences and the number of second homes. These series are used to calculate the populations of *communes* that count under 10,000 inhabitants.

The DGFiP also issues the *communes* with information on of the number of tax rolls dispatched and on the taxable base. Their response, together with the response from taxpayers, is instrumental in improving this source data. The fact is that the net result of any omissions would be lower tax revenue, and so the *commune* has every reason to notify the DGFiP of any omissions identified. Similarly, since the net result of any overage (premises mistakenly included in the listing) would mean the taxpayer ends up being overtaxed, overage is generally also reported to the DGFiP.

The INSEE systematically runs a profiling analyses on the series put together, and if the analysis surfaces any conflicting trend (such as a sharp drop followed by a sharp rise), the DGFiP is queried to detect any processual artefacts. Any artefacts confirmed by the DGFiP are corrected.

Assessments on the quality offered by this datasource indicate that even if the file is not compiled according to the same design concepts as the census, it does yield reliable indications on housing stock trends, and it is precisely the trend factor – and not the level factor – that is integrated when calculating population.

e) Conceptual shifts between the 1999 general census and the new census scheme

These conceptual shifts can influence the analysis of trends in a *commune* population. For instance, the fact that the new census scheme counts residential-school students over the age of majority in the *commune* where their school is located rather than the *commune* where their parents live has a strong impact on the town population of the residential-school *commune*.
Those *communes* initially expected to be strongly impacted by this kind of drift should therefore be listed in order to properly validate the figures for these *communes* and, by extension, the statistical output that is based on these *communes* (or the zonal areas of coverage). This is an effect that users of this data need to be alerted to.

f) Conclusive validation

Given that each component input is validated progressively as it is brought in, there is not, strictly speaking, a final validation step. It is at this juncture that the full set of qualitative and quantitative inputs substantively explaining the population figure can be brought together. This is pivotal if the INSEE is to be in a position to respond to any queries coming in from the *communes* or from other end-users of the population figures.

8.2. Validation of detailed statistical output

Validation extends beyond the basic population figure to cover all detailed statistical output. The quality validation on the raw final figure does guarantee that the vast majority of quality-related issues (coverage, effects tied to the sampling system, etc.) have been detected and catered for. However, quality is also backed up by two further validations.

Validation of the questionnaire coding system

The coding step starts with an automated text coding phase (matching to established classifications for activity status and profession) followed by an expert coding phase where noncoded cases are channelled on to operatives at regional centres The INSEE is currently deploying a quality control system to cover this process. The working principle is to run a second coding operation on a subsample of schedules processed though both coding modes (auto-coding and expert coding), and examining any drift between the two.

The objective is to measure the quality performance of both automated coding and manual reprocessing, starting with a global percentage of correctly-coded questionnaires and then honing in towards percentage figures for broad subgroup variables, such as the percentage of respondents correctly-coded as executive managers).

These quality metrics and control measures can also be exploited to develop on the learning files used for automated coding in order to improve performance, refocus training for the operatives tasked with reprocessing the responses left unmatched following automated coding, and ultimately ensure a good level of quality management. This step was first implemented in 2006, and the benefits were already tangible from the outset of the 2007 campaign.

Reasonableness checks on detailed counts

Detailed statistical output is analyzed via univariate ('flat') tabulation (counts on the occurrences of each different variable: age-sex pyramid, marital status, family structure, employment, housing stock, and so on). The process involves cross-checking the output against results from other sources, and assessing robustness by analyzing patterns of change between surveys. Level-scale analysis can then give an evaluation of the level of quality achieved through the adjustment procedures. Robustness analysis coupled with estimates of accuracy needs to be able to define the target level of granularity required for the information earmarked for release (to resolve issues like whether to release 5-year or 10-year statistics on ages). The validation stage is also the ideal point at which to sketch out and if possible quantify the effect of changing questionnaires. This is crucial data for end-users seeking to analyze patterns of change in relation to the previous census.
The analyses conducted over the first few years have surfaced a handful of adverse effects created by the adjustment procedure (effects that we minimized in later surveys) or by over-fuzzy data entry protocols (an error corrected from 2005 on). They also prompted analysis on the effects of changing certain questions (as highlighted through the questions on employment and family structure).
We currently use a 'screen' to analyze trends in the core statistical variables, and the regional census offices are invited to look into cases where these trends pose problems.

9. Future perspectives for analyses into census coverage.

We have outlined above how a significant in-process control effort is led iteratively, from census data collection and over the three months following questionnaire check-in. These checks prompt immediate correction of any problematic personal schedules and/or housing schedules.

On a more global level, it is equally important to ensure that the census has not counted the same people several times over while others slip through the net. There are several formats of census coverage analysis:

1. An overcount analysis, which hinges on detecting people censused twice over. Overcount analysis is based on the permanent demographic sample (*EDP*) – a sample covering roughly

1% of the French population and that is regularly updated with civil registration records and census form data. The procedure governing inputs to this *EDP* sample can identify whether several schedules have been put together for the same person. Given the census methodology now deployed, questionnaires should be compiled from five successive campaigns in order to draw solid conclusions. Particular focus is given to producing net census overcount figures, before moving on to characterize the populations concerned in terms of residency profile (larger or smaller *communes*, standard households or community housing) or age and gender. The census-taking methodology, and most importantly the rate of EDP coverage, means that the results of this analysis will only carry national-level significance.

2. Omissions analysis, which could draw on administrative source files related to universal welfare healthcare coverage – as under the French “Assurance-Maladie” scheme – or tax returns.

These avenues open up more promising perspectives than those offered by the kind of post-enumeration surveys that were led in France in 1962 and 1990. The evidence suggests that it is more relevant to quality-control a survey by cross-comparing the data against administrative source files rather than against other surveys that are liable to present the same kind of flaws (especially the bias introduced by difficulties with reaching households for interview).

These investigations, which are centred on the data collection process, will further complete the *RIL* quality control checks described in section 8.1.