UNITED NATIONS **E**

Economic and Social Council

Distr.
GENERAL

ECE/CES/AC.6/2008/8
4 March 2008

ENGLISH
Original: FRENCH

ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Meeting on Population
and Housing Censuses

Eleventh meeting
Geneva, 13-15 May 2008
Item 4 of the provisional agenda

DATA EDITING AND VALIDATION

VALIDATION OF CENSUS DATA IN FRANCE

Note by the National Institute for Statistics and Economic Studies, France

SUMMARY

This document describes the procedures developed by the National Institute for Statistics and Economic Studies for the validation of census data in France. It sets out the method adopted for calculating the population of a given district, followed by the validation procedures applied to all the data sources used in calculating the population of a district.

GE.08-21605 (E) 170408 220408

I. METHOD OF CALCULATING POPULATION FIGURES AND STATISTICAL RESULTS

1. There are two requirements imposed on the National Institute for Statistics and Economic Studies (INSEE) with regard to the publication of population figures:

(a) It must publish the population of all districts every year;

(b) To ensure that all districts get equal treatment and that regional populations are not disadvantaged by inconsistency of cumulative totals, population figures should relate to the same year for everybody. It would not be acceptable to use the 2004 figure for one district and the 2008 figure for another.

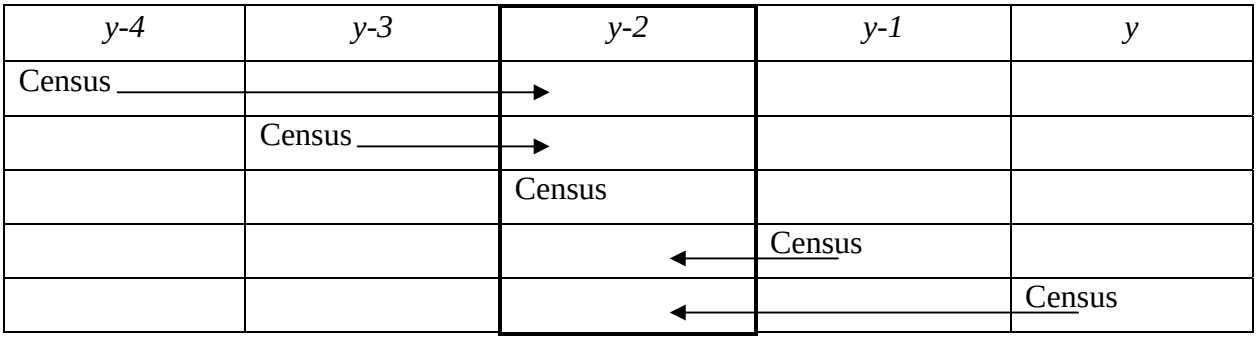
2. INSEE calculates these populations as set out below.

A. Household populations

3. The method used for districts of 10,000 inhabitants and above relies on a rolling average based on the samples over five years. From the aggregate of the five samples from year *y-4* to year *y*, an average population per dwelling is calculated, which is representative of the situation at mid-period (year *y-2*). This ratio is then multiplied by the number of dwellings at the start of *y-2*, as given in the located buildings register (RIL), to obtain the population of the households in the district.

4. For districts of fewer than 10,000 inhabitants, meanwhile, the population figure has to refer to the median year of the quinquennial cycle in order to be consistent with districts having 10,000 inhabitants or more.

5. The data are supplied by censuses, in accordance with the following diagram.



6. For districts surveyed in *y-2*, the results are saved.

7. For districts surveyed in *y-1* and *y*, the *y-2* population is obtained by interpolation between the census survey and the last result published. For the districts surveyed in *y-4* and *y-3*, the calculation involves extrapolation between the results of the census survey and *y-2*: this is based on local housing-tax data, which indicate any change in the number of dwellings per district, and is adjusted to take account of the differential between the change in the number of dwellings and the change in the number of inhabitants. The differential measured between the last two censuses is applied to the change measured by the housing-tax data to give the change in population.

B. Non-household population

8. The population figures for institutions ("communautés") were collected on 1 January 2006, the same date as those for the household population. Where a district's institutions were surveyed in 2006, the figures arrived at are saved. Where they were surveyed in 2004 or 2005, their population is updated by adding the population of new institutions and subtracting that of institutions that no longer exist. These changes are made on the basis of the institutions register. If they were surveyed in 2007 or 2008, the population of institutions is brought up to date by interpolation between the last result published and the result provided by the census of institutions.

9. Populations of homeless persons, those living in mobile homes or those whose principal residence is a hotel are not updated. They are treated as being unchanged over the four years following the year of the census. The new figures replace the earlier ones as they become available.

II. VALIDATION PROCEDURES

10. The following elements are therefore necessary for calculating a district's population:

(a) Files containing data from the various census surveys;

(b) Housing stock drawn from the RIL;

(c) Institutions registers;

(d) Housing tax.

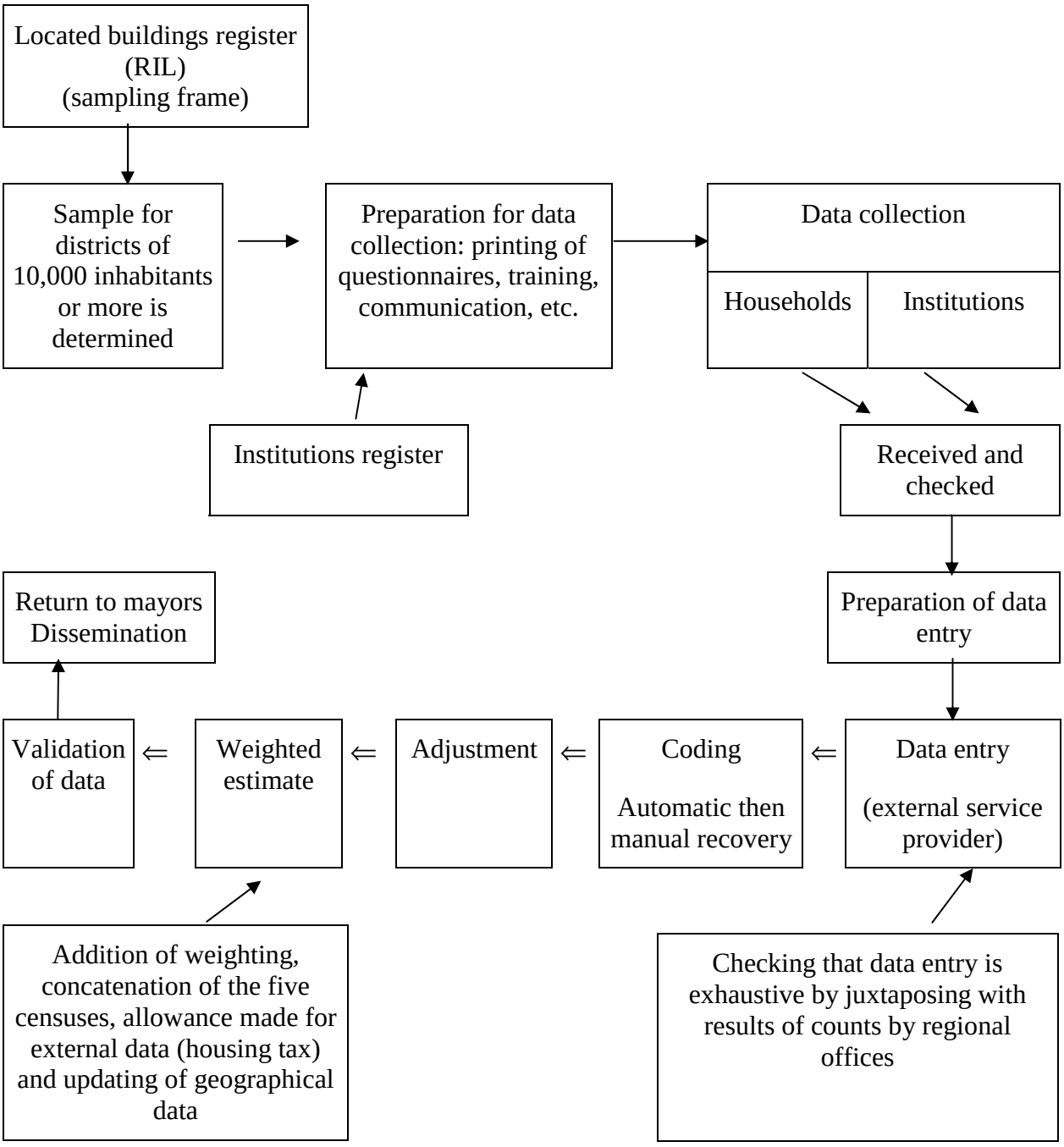
11. Each of these elements will be validated. Validation takes place in the course of the preparation or processing of these elements, since errors are much easier to correct where they are detected in good time.

12. Validation also includes an examination of two factors that may affect the quality or the interpretation of data: changes arising out of differences in perception from one census to the next and the effects of adjustments. In large districts, validation will also, given the need for weighting and extrapolating the census result, take account of the effect of "large addresses", which can generate cluster effects.

13. Validation has two objectives: first, to ensure the statistical quality of the data to be issued and, second, to understand their development, assess their robustness and use them as a basis for employment concerns. This element is particularly important if differences in perception particularly affect a given district.

14. The flow chart below shows the sequence of the various stages of the census.

Chart. Census progression



A. Housing stock from RIL

15. The sampling frame in each large district is the relevant RIL, which is a list of all buildings, whether residential, administrative, industrial or commercial premises, identified and located by their address, using a geographical information system.

16. The RIL was originally compiled on the basis of the 1999 general census and has been kept up to date since then through the use of administrative data, such as building permits and local tax records, or postal records, such as Post Office address lists. It is checked by the districts every year before final validation by INSEE.

17. An RIL is thus validated extensively just before being put to use as a sampling frame.

18. The regional offices then issue a report for each district, analysing the quality of the register and evaluating the district's participation in assessment operations and the sources used for updating.

19. Once an RIL is validated, INSEE proceeds to make a range of supplementary information available to regional offices concerning housing number series, large addresses and discrepancies between data collected and the RIL, in the year following the delivery so as to enable them to gain a better understanding of the statistical impact of changes to the RIL.

20. Quality measurements of RILs are conducted annually in the field at both national and regional level by INSEE investigators.

21. The objective is to measure the completeness of an RIL as regards a shortfall or surplus of addresses in the residential and associated housing category. A quality measurement aims to ensure that all addresses actually existing are properly registered in the RIL and, conversely, that all the addresses in the RIL do in fact exist.

Table 1

Shortfalls and surpluses of residential addresses in the RIL

		2003	2004	2005	2006	2007
Addresses	Shortfall	2.3%	1.9%	1.3%	1.3%	1.4%
	Surplus	2.0%	1.9%	1.8%	1.7%	1.4%
	Balance surplus/shortfall	-0.3%	0%	0.5%	0.4%	0%
Dwellings	Shortfall	2.3%	1.9%	1.1%	0.9%	1%
	Surplus	2.1%	2.3%	1.8%	1.4%	1%
	Balance surplus/shortfall	-0.2%	0.4%	0.7%	0.5%	0%

Source: INSEE, RIL quality surveys, 2003-2007.

B. Files containing data collected

22. The quality of such files depends on two factors: the quality of the data collection and the quality of the data entry. Adjustments are then made with a view to improving the quality by correcting non-responses or inconsistent responses. The coding stage also undergoes quality control.

23. Data collection is often the weakest link in the statistical process. Its quality depends on precise and rigorous procedures but also on the support of respondents (which determines the sincerity of their responses) and the commitment of those gathering the information, namely the enumerators. Their quality is ensured by rigorous preparation, involving skills and training, and by monitoring. Checks are carried out by the districts and INSEE during data collection. After the collection, checking is the responsibility of INSEE, which can then verify the data collected and, where necessary, correct them.

24. For principal residences that it has not been possible to survey (where occupants cannot be contacted, are away for a long period or refuse to reply), the enumerator completes a non-surveyed dwelling form giving the presumed number of occupants, which helps in adjusting the population figures. Risks of omission are thereby reduced.

After data collection: checks conducted by INSEE

25. Once INSEE has received the documents transmitted by the districts, checks are conducted first in the office and then in the field. The checking procedure is appropriate for the new way of conducting the census. By contrast with general censuses, more detailed checks can be made for each district, owing to the smaller volume of data collected every year and the smaller number of districts involved. Housing-tax files play an important role in office checks and make it possible to verify the completeness of data relating to a given district or collection area. Such checks, whether in the office or in the field, are conducted as close as possible to the collection area. The errors thus picked up are corrected.

26. The checks are selective in that they pass through a succession of filters:

(a) First, checks carried out as soon as questionnaires are received and registered make it possible to assess quality for each district. The checks consist primarily of counting the returns by reading the bar codes appearing on each return. The resulting figures are compared both with those established by the district on the conclusion of the data collection and with the expected number of returns, based on the previous census and observed trends. Access to these figures is also available within the district, to census enumerators, which means that, where necessary, targeted checks can be made of consistency and quality;

(b) In about 10 per cent of cases, this assessment leads to detailed checks at the office, focusing on all or part of the district concerned;

(c) Lastly, cases not resolved in the office undergo checks conducted on the ground by INSEE (some 28,000 such checks were carried out in 2007). One dwelling in every 200 undergoes checks on the ground. In most cases, these checks confirm the data collected but they also make it possible to detect errors. These are rectified and the faulty returns are replaced by error-free documents, which can thus form part of the data entry series.

27. A particular effort is made with regard to the number and quality of non-surveyed dwelling forms. The total number of such forms needs to be reduced by several percentage points so that the estimates are not skewed. The rate is currently around 3 per cent, thanks to strenuous efforts made in the course of household data collection.

28. It is also important to ensure that non-surveyed dwelling forms are properly filled out for principal residences and that the stated number of occupants provides a good estimate of household size. Checks on the ground make it possible to verify such information.

Data entry validation

29. Data entry is carried out by an external service provider. Data are entered in optical form from the questionnaires, which are scanned and then submitted to pattern-recognition software. The process is completed by the entry of the data concerned using ordinary language. The service provider is given precise specifications as to what is expected in terms of quality, as regards security in transit and on the service provider's premises - physical and data processing security must be guaranteed and an undertaking given by everyone involved not to divulge the data processed - and the quality of the data, as indicated by the number of documents entered per collection area and the maximum error rate by variable. Lastly, the documents are counted and discrepancies with the first bar-code count conducted at INSEE (see above) are checked.

30. The procedure thus involves the documents (individual census returns and housing forms) being counted three times: once by the district in the course of conducting quality controls, then by INSEE at the monitoring stage and finally by the data-entry service provider. The comparison of these three counts results in extremely high quality, considering the number of returns involved.

31. Lastly, data entry quality is measured by double entry (see box).

Double data entry to verify variable data entry quality

Data entry errors are assessed by means of double data entry carried out on a sample of forms by another service provider, which, working from the scanned images, uses the same data entry rules. Any discrepancies are analysed at INSEE in order to establish error rates on the coding of each variable. Double data entry is carried out during the data entry process, which means that, where necessary, data entry protocols can be corrected for subsequent batches. Regular contacts and meetings take place with the service provider in order to anticipate problems or resolve them as fast as possible. The service provider is requested to make adjustments or improvements during the census period or, where more substantial improvements are required, between censuses, on the basis of information from the double data entry. For example, the improvement of pattern-recognition tools has contributed to a significant decrease in the error rate of data entry of dates of birth. In that connection, it should be noted that, even for very simple data, such as dates of birth, it is unrealistic to expect a zero error rate: since information is written by hand at the outset, errors of interpretation are inevitable.

32. **Adjustments**. The aim is to improve the quality of files containing data collected. This is done by analysing total or partial non-responses, which are then imputed, but also by analysing census returns that contain discrepancies, such as a 95-year-old person claiming to be in work. The "hot-deck" procedure is often used in such cases, whereby the missing information is supplied by using the response of a "donor" having similar characteristics. In such cases, too, adjustments are refined year by year on the basis of previous results. The end result is an effective quality improvement system.

33. In particular, the adjustment of non-surveyed dwelling forms makes it possible to calculate the population of a district by imputing the population of principal residences in which contact with the enumerator did not occur, owing to extended absences, the impossibility of contacting a particular respondent or refusal to cooperate. For principal residences where the number of occupants is provided (more than 80 per cent), the number of individuals in the dwelling is still "imputed" from the household figures given in the non-surveyed dwelling form. For other residences, an imputation is also made, following a method that reproduces the pattern, by household size, observed as the national average for residences on which data exist. An adjustment is made to cater for the fact that a small number of forms are incorrectly completed for non-principal residences, so as to ensure that the number of principal residences in a given district is not exceeded.

C. Institutions register

34. Like the RIL, the institutions register was established at the time of the 1999 census and is kept updated using administrative sources, such as files from health or social facilities, boarding schools, prisons or juvenile reform institutions. Like the RIL, it is transmitted to districts for appraisal just before a census. The districts are sent a list of their institutions and asked to indicate which have disappeared, which INSEE might be unaware of and which might have changed their address. INSEE then approves proposals by the local authorities to add institutions to the register or to remove them.

35. Once data collection commences, the enumerator in charge of institutions is requested to conduct a final validation of the institutions list.

36. Lastly, INSEE conducts an annual update of the RIL and the institutions register. This involves establishing the geographical location of each institution listed in the register, and the RIL is then used as a reference for their addresses and their geographical characteristics, on the basis of their X and Y coordinates and their location in a given area, canton or subdistrict. Second, it involves cross-checking for duplications or omissions between the two registers, such as may occur where an institution has changed into a set of ordinary dwellings, or vice versa.

D. Housing-tax files

37. INSEE annually receives from the Directorate-General for Taxation a file containing a list of all premises subject to housing tax. The first action is to delete premises that do not qualify as dwellings, such as garages or car parks.

38. The file is then used for the following two purposes:

(a) To follow up and monitor data collection for the district concerned;

(b) To establish total numbers for all dwellings, for principal residences and for secondary residences. These figures are used to calculate the population of districts with fewer than 10,000 inhabitants.

39. The districts also receive information from the Directorate-General on the number of tax rolls issued and on taxable income. Naturally, since any omission results in lower tax revenues, the local authorities inform the Directorate-General of any such omission. Since any surplus - as a result of premises mistakenly appearing on the list - leads to the taxpayer concerned paying additional tax, surpluses are also reported to the Directorate-General.

40. INSEE systematically analyses the profile of the dwellings thus constituted and, in the event of any atypical development, such as a sharp fall followed by a sharp rise, requests the Directorate-General to identify any management artefacts. Artefacts confirmed by the Directorate-General are corrected.

E. Extrapolations in large districts

41. The population of large districts is obtained by extrapolating the sample accumulated over five successive years of data collection. The validation of the figure is contingent on the robustness of the indicators obtained by analysing each annual survey and each cumulative figure. The observed trends represent an empirical indicator of how reliable the overall estimate is. Other criteria are also taken into account for the validation: the consistency of the estimate with the results of the previous census and indicators arising from external factors also provide indications of the reliability of the figure. Districts for which the data are considered (by INSEE or the regional offices) to be the most surprising are subjected to detailed analysis aimed at uncovering any defects in the RIL or cluster effects that have not been dealt with. Districts affected by such a problem may undergo a separate assessment aimed at correcting such flaws.

42. The attached form was used for the validation of provisional estimates based on four censuses. In addition, an analysis was conducted at subdistrict level to highlight areas whose unusual development would have a particular impact on the final result.

F. Perception gaps between the two censuses

43. Perception gaps can affect the analysis of trends in the population of a given district. It is therefore worth compiling a list of districts that are likely to be particularly affected by a perception gap in order to cast some light on the validation of such districts and statistical analyses based on them or on the surrounding areas. For example, in the latest census boarding-school pupils who had reached the age of majority were counted in the district of their boarding school rather than in their parents' district, as a result of which the local population of a district where a school was located was significantly higher. Such an effect should be drawn to the attention of data validators and users.

G. Final validation

44. Since each factor is validated as and when appropriate, there is no final validation stage, strictly speaking. It is simply the point at which all the qualitative and quantitative factors that may explain population figures are assembled. It provides essential material for replying to any question that may be put by district authorities and other users of population figures.

III. VALIDATION OF DETAILED DATA

45. Validation applies not only to population figures alone but to all detailed data. Of course, the validation of the total figures ensures that most of the problems relating to quality, such as coverage or sampling effects, have been detected and taken into account. Two further validations are then undertaken.

A. Validation of questionnaire coding

46. Coding involves first an automatic phase in accordance with such classifications as activities or occupation. Operators in regional offices then process uncoded cases. A system for controlling the quality of this procedure is currently being developed. A second codification of a sample of forms representing the two codification methods - automatic and manual recovery - is then undertaken and the discrepancies are analysed.

47. The objective is to measure the quality of automatic coding and of manual recovery, by first identifying an overall percentage of well-coded cases and, in due course, percentages representing a wide range of variables, such as the percentage of well-coded senior executives. A second objective of the codification quality control system is to measure what proportion of individual forms cannot be coded, because they use vague terms such as “civil servant” without any other description. This will make it possible to establish a realistic target for quality and avoid over-punctilious scrutiny.

48. This measurement and control system also makes it possible to enhance the instructions for the use of automatic coding. This will lead to improved performance, more focused training for manual recovery coders and, lastly, good quality-management. The contribution made by this process will have an impact on any subsequent census.

B. Probability control on detailed data

49. Detailed data are analysed on the basis of the simple tabulation of a range of variables, such as the population pyramid, marital status, family structure, employment or housing stock. The data are compared with data from other sources and their robustness is assessed by means of an analysis of trends from one census to the next. Through analysis of the various levels, it is possible to assess the quality of the adjustment procedures. In combination with accuracy tests, these robustness assessments make it possible to determine the level of detail required; for example, should ages be given quinquennially or decennially? Validation also provides the

opportunity to determine and quantify, where possible, the effect of changing questionnaires. This is essential information for users who wish to analyse developments since a previous census.

50. Analyses carried out on the first years made it possible to identify a number of undesirable consequences of the adjustment procedure (some of which were eliminated in subsequent censuses) and data entry protocols that were too imprecise (that defect has been rectified since 2005). In turn, those analyses prompted analyses of the consequences of changing the questions on such topics as employment and family structure.

51. The current procedure is for trends in the main statistical variables to be sifted and analysed, after which regional offices are invited to analyse problematic cases.
