Proceedings of Statistics Canada's Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

REDESIGN OF THE FRENCH CENSUS OF POPULATION

Jean-Michel Durr,[1] Jean Dumais[2]

ABSTRACT

Census-taking by traditional methods is becoming more difficult. The possibility of cross-linking administrative files provides an attractive alternative to conducting periodic censuses (Laihonen, 2000; Borchsenius, 2000). This was proposed in a recent article by Nathan (2001). INSEE's redesign is based on the idea of a "continuous census," originally suggested by Kish (1981, 1990) and Horvitz (1986). A first approach that would be feasible in France can be found in Deville and Jacod (1996). This article reviews methodological developments since INSEE started its population census redesign program.

KEY WORDS: balanced sampling, census, continuous census, calibration

1. INTRODUCTION

1.1. Reasons for the Redesign

France has been conducting censuses for many years to measure the de jure population of its administrative districts and to describe the socio-demographic characteristics of its territory at all levels of geography, from districts of communes to the country as a whole. The 1999 census was conducted in the usual manner. For various reasons, however, we decided to re-examine the census. First, because of budget problems, the period between censuses is getting longer. Before the war, censuses were taken every five years; then the gap grew to seven years, then eight, and finally nine years (between the 1990 and 1999 censuses). Moreover, the public does not always understand the need for such a massive operation at a time when the number of administrative files is increasing, even though that same public has expressed serious concerns about the cross-referencing of such files. In addition, the decentralization that has been going on in France for over 20 years has generated numerous requirements for statistics in support of local policy-making. As the supreme source of local information, the census had to adapt to these changes and provide fresher yet still highly detailed data.

As a result, a population census redesign program was established at INSEE in the late 1990s. Since France has no population register and, in view of the circumstances, is unlikely to institute one, the decision was made to consider a compromise solution that would combine annual sample surveys with the use of non-nominative administrative files that INSEE is authorized to use solely for statistical purposes. Communes whose population is below a certain threshold (10,000 for the moment) will be covered by annual take-all surveys with a rotation period of five years. For the other communes, a sample survey will be conducted each year, with the entirety of the commune being covered within the same five-year rotation period. To carry out this redesign, a new legal framework was needed. The project was submitted to the Conseil d'État, which recommended on July 2, 1998, that the government table draft legislation in Parliament. Aside from the need to provide the census with a legal basis, the Conseil was of the view that since population counts were referred to in over 200 statutes or regulations, making a major change in the way they were produced would

[1] Jean-Michel Durr, population census redesign program, INSEE, Direction générale, 18 boul. Adolphe Pinard, 75675 Paris CEDEX 14, France.
[2] Jean Dumais, population census redesign program, INSEE, Rhône Alpes, 165 Garibaldi, 69401 Lyon CEDEX 3, France.

require legislative approval. Within this framework, the purpose of the legislation was essentially to set out the principles and rules governing the organization of the census.

The operation was placed under State responsibility and control: INSEE was to establish the collection framework (concepts, protocols), select the samples, ensure the quality of the information collected, and process and disseminate the data. The communes or commune groups were to prepare and conduct the census surveys. The State would provide financial assistance to cover the costs.

1.2. Quality Goals

The program has the following **quality goals**:

1.2.1. Data quality

Timeliness: The goal is to be able to disseminate by the end of year A the de jure population of all administrative districts as at January 1 of year A-2; a statistical description of all geographic units (communes and commune groups, districts of major cities, lands, etc.) as at January 1 of year A-2; and a statistical description of France and its major geographic units (regions, etc.) as at January 1 of year A. In comparison with the general census, the redesigned census will produce similar population and housing data an average of three to four years earlier.

Relevance: The data produced must be relevant to local needs. In particular, data that are worth studying only at levels of geography far above the commune will be set aside in favour of data that are more useful for local purposes. What data will be collected will be determined by the Conseil national de l'information statistique (CNIS), whose membership includes representatives of various categories of producers and users of public statistics. A CNIS working group has proposed changes while at the same time preserving the necessary continuity with previous censuses and limiting the response burden.

Precision: The census must provide data that are meaningful for all levels of geography in France. The data produced must be sufficiently precise, even at the subcommunal levels, for the most useful cross-tabulations at those levels. This means, in particular, distributions by sex and age, by type of activity and socio-professional category, and by type of housing. It must be possible to estimate the precision of the data, and users must be informed of that precision.

User-friendliness: To avoid annoying users, the data produced must be easy to understand and comparable in use to data produced by a general census.

1.2.2. Process quality

Response burden: To limit the response burden for the public, the amount of information collected must be kept to a bare minimum. In particular, information available for the same level of geography from other sources will not be collected in the census unless it can be used to produce useful cross-tabulations with other variables. As in previous censuses, the personal questionnaire will be confined to one double-sided sheet of paper.

Questionnaire: Since collection is by the drop-off/pick-up method, the questionnaires must be universally accessible. To ensure that the questions will be understood, qualitative testing was conducted using focus groups. In addition, a collection test was carried out on 4,000 dwellings in the first half of 2001.

Confidentiality: Data gathered in the census are protected by law. Personal information collected in the census can be accessed only by authorized persons. The data are for INSEE and can be used only for statistical purposes. Only data essential to the preparation and conduct of census surveys are shared with communes or commune groups, on a need-to-know basis.

Technical and organizational robustness: Because of the volume of data processed and the importance of the census, the program must be based on tried and true technical innovations. Furthermore, the robustness of the census apparatus must be evident in the launch of the operation. Technical or functional innovations can be introduced at any time in the census cycle as part of evolving maintenance or specific projects. The annual surveys can be used to test the effectiveness of such projects before they are applied to the entire process. However, major changes such as questionnaire updates will generally be made only for the beginning of a five-year cycle. The organization of the census will depend on a balanced partnership between INSEE and the communes. INSEE must be capable of building the proposed structure within its budget and its work program by reorganizing its operations. Similarly, the communes and intercommunal cooperation bodies must be able to support the census organization. The yearly cycle of surveying large communes and the option that small communes will have of delegating collection to an intercommunal body are likely to promote the professionalization of collection workers.

With the integration of census operations into the annual work program of the regional offices, and the fact that the operation is one-seventh the size of the general census, INSEE will have tighter control of the census. Instead of having 110,000 census agents collecting data from 60 million people in 36,700 communes in a particular year, it will have only 18,000 agents visiting roughly 9 million residents in about 8,000 communes.

The division of responsibilities between INSEE and the communes, the resources that the communes will require, and the validation processes for the various stages will be set out in a decree.

Cost control: With the five-year collection cycle, the financial burden of conducting the census can be spread over a longer period. For communes with a population of more than 10,000, the cost of the redesigned census will be lower than the cost of the current census of population. On the other hand, for communes with fewer than 10,000 residents, the cost should be equal to that of a general census, but it would be every five years instead of the roughly eight-year cycle of the general census. The cost of the redesigned continuous census will probably be less than 30.5 million euros (2000 euros) a year. However, a slightly larger budget in the first few years would help to iron the kinks out of the collection process.

2. SAMPLING STRATEGY

The commune is the linchpin of the redesign effort. The set of “small and medium-sized communes” (those with a population of less than 10,000) will be sampled at an average rate of 20% a year, and all their dwellings will be visited; all “large communes” will be visited annually, but only a fraction of their dwellings will be surveyed.

2.1. Small and Medium-sized Communes

Let’s start with “small and medium-sized communes”. In each region, five rotation groups of communes will be formed using data from the 1999 population census. They will consist of balanced samples (Deville and Tillé, 1999, 2000) of the age-sex distribution of the communes’ population. This approach should help minimize year-to-year variation due to sampling.

Figures 1 and 2 show how balanced the five rotation groups will be. They contain box-and-whisker diagrams of two variables measured in the 2,811 small communes in Rhône-Alpes in the 1990 population census. For each rotation group, both the quartiles and the range of the distribution are shown. It is interesting to note how similar the charts are. The “number of women aged 20 to 39” variable was used to form the groups. Neither the number of principal residences nor any of the household or dwelling variables plays a part in the balancing.

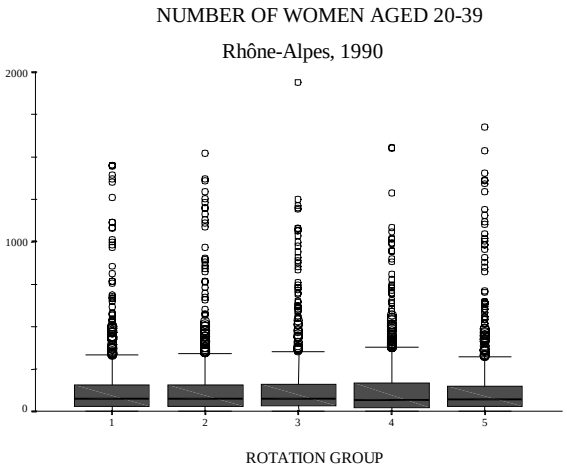

Figure 1

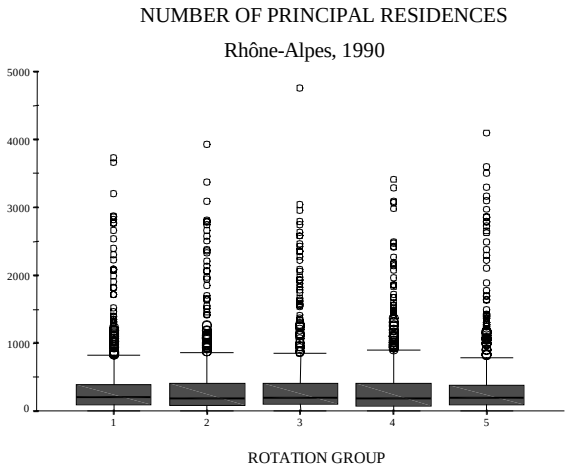

Figure 2

Each year, the population and housing in all the communes in one rotation group will be fully enumerated. Hence, each "small commune" will be completely enumerated once every five years, and a fifth of all the "small communes" will be covered each year.

2.2. Large Communes

The "large commune" sample will be based on the "répertoire d'immeubles localisés" (RIL) (inventory of located buildings). The RIL is a list of buildings (residential, institutional or commercial) identified individually so as to generate a digitized map. Initially, the RIL will be populated with data from the 1999 census, which will provide a statistical portrait of each residential building.[3]

The RIL will be continually updated using building permits, demolition permits, utility records (water, gas, hydro, etc.), information supplied by local governments, and field observations. Thus, the RIL may be used to create a building sample frame for "large communes".

In each IRIS2000[4] of each "large commune", five rotation groups of addresses will be formed using the same sampling model as in "small communes". Three additional strata will be created in each IRIS2000: one for industrial buildings (plants, warehouses, etc.), another for collective dwellings (institutions, group homes, communal groups, boarding schools, etc.) and a third for new addresses.

One fifth of the industrial buildings will be visited each year to verify that they contain no dwellings (custodian's quarters or space converted for habitation); any dwellings found in such buildings will be considered self-representing because of their special nature. All collective dwellings will be covered each year; 20% of them will be visited, and the population counts of the remaining 80% may be updated by telephone. Finally, all new residential buildings will be enumerated so that they can be placed in a rotation group and their statistical profile can be recorded.

As noted above, each address rotation group will be visited once in each five-year period. Data will be collected in the rotation groups by one of the following methods:

- Each year, a list of the dwellings at the residential addresses in the current year's rotation group will be prepared, and a 40% sample of those dwellings will be invited to take part in the census for that year.

[3] In the 1999 census, a building is defined as the set of dwellings served by the same staircase; thus, a single physical building can consist of more than one "census building".

[4] An IRIS2000 is a set of "îlots regroupés selon des indicateurs statistiques" (blocks grouped by statistical indicators), a homogeneous area with a population of about 2,000.

- A subsample consisting of half the addresses will be drawn from the list of addresses in the year's rotation group; in other words, 50% of the dwellings will be sampled on average.

"Large communes" may request that the sample be increased to as much as 100% of the rotation group's dwellings.

In summary, the annual sample will consist of some 8 million individual forms, 6 million from "small communes" and 2 million from "large communes".

3. OVERALL AND DETAILED ESTIMATES

In the continuous census system, three sets of estimates will be produced and published each year: a set of de jure population estimates, a set of detailed estimates (from which the de jure population estimates will be derived) and a set of overall estimates that will be used to calibrate the detailed and de jure population estimates.

3.1. Overall Estimates

According to current dissemination plans, the national and regional results of the survey conducted at the beginning of year A will be published on December 31 of year A. These estimates will be the overall estimates for year A. In addition, the results for each "small commune" visited during the year A collection campaign will be published on the same date.

3.2. Detailed Estimates

Administrative files will supply additional information at a sufficient level of detail. It will then be possible to measure the systematic error between what has been observed and what is in the files for similar objects (buildings, blocks, etc.). This systematic error in carefully chosen aggregates can be used to produce an adjustment factor which will then be applied to the administrative data to ensure that their adjusted totals match the census estimates.

Current plans are to use administrative files at a level of geographic aggregation (building, block, census agent district, etc.) that will provide information about individuals (age and sex according to health insurance files) or their dwellings (property tax files).

Detailed results for year A-2 will be released on December 31 of year A.[5] These detailed results will be a blend of survey data (large communes) or census data (small communes) with synthetic data.

The synthetic data will be obtained from the relationship between observed data and administrative data for the same point in time and space. For example, for commune C of Group II enumerated in year A-3 (census count denoted $R_{C,II}^{A-3}$), the imputed census count for target year A-2 will be given by

$$\tilde{R}_{C,II}^{A-2} = R_{C,II}^{A-3} \times \frac{Adm_{II}^{A-2}}{Adm_{II}^{A-3}} = R_{C,II}^{A-3} \times \frac{\sum_{c \in II} Adm_c^{A-2}}{\sum_{c \in II} Adm_c^{A-3}},$$

where Adm_c^a is the value derived from administrative sources for commune c and year a.

In the continuous census, for a "small commune" surveyed in years A-5 and A (see the table below), person variables (age, sex, labour force activity, occupation, etc.) and dwelling variables (household size, number of rooms, tenure, conveniences, etc.) will be measured at two points in time.

[5] Acquisition and processing of administrative files are expected to take about two years.

	A-6		A-5		A-4		A-3		A-2		A-1		A	
Gr I		Adm		Adm	R	Adm		Adm	?	Adm		Adm		Adm
Gr II		Adm		Adm		Adm	R_{II}^{A-3}	Adm	$\widetilde{R}_{C,II}^{A-2}$	Adm		Adm		Adm
Gr III		Adm		Adm		Adm		Adm	R_{III}^{A-2}	Adm		Adm		Adm
Gr IV	R	Adm		Adm		Adm		Adm	?	Adm	R	Adm		Adm
Gr V		Adm	R	Adm		Adm		Adm	?	Adm		Adm	R	
Total	5R	∑Adm	5R	∑Adm	5R	∑Adm	5R	∑Adm	5R	∑Adm	5R	∑Adm	5R	∑Adm

In addition, for Groups IV and V, the synthetic estimates for year A-2 could benefit from the information collected in the campaigns for years A-1 and A respectively. Adjustment factors could be computed in relation to the most recent census and used to produce backward projections for the intercensal period. For example, for commune D in Group IV, we can compute the following:

$$\Theta_1 = R_{D,IV}^{A-6} \times \frac{\sum_{c \in IV} Adm_c^{A-2}}{\sum_{c \in IV} Adm_c^{A-6}} \text{ and } \Theta_2 = R_{D,IV}^{A-1} \times \frac{\sum_{c \in IV} Adm_c^{A-2}}{\sum_{c \in IV} Adm_c^{A-1}}.$$

It is virtually certain that these two series, extrapolations and backward projections, will not match. Nevertheless, it is best to publish just one set of estimates for any area and any point in time. It makes sense to produce a "composite" series whose end points are tied to census data. The following linear combination may accomplish just that while giving more weight to the more recent survey data:

$$\widetilde{R}_{D,IV}^{A-2} = 0.2 \times \Theta_1 + 0.8 \times \Theta_2 .$$

Similarly, for commune E in Group V, with Θ_1 and Θ_2 appropriately defined, we would have:

$$\widetilde{R}_{E,V}^{A-2} = 0.4 \times \Theta_1 + 0.6 \times \Theta_2 .$$

Adjustment factors Θ will have to be calculated for relatively detailed population strata, such as age-sex classes, so as to keep as much demographic and geographic flexibility as possible in the census adjustment. The quality of the administrative files and local disparities will dictate the level at which the adjustment can be made most conveniently. The same process can be applied to large communes if we replace "small commune" with "address".

Finally, when every commune in every group has been imputed, the estimated total for a variable of interest from the imputed file (detailed estimates) is unlikely to match the total estimated from observations alone (overall estimates published two years earlier). It has therefore been decided that the detailed estimates will be calibrated on the overall estimates. Once again, the level of calibration will depend on local trends and the quality of the overall estimates.

3.3. De Jure Population Estimates

The de jure population estimates are the third set of estimates derived from the census. They are the population figures that are used, by law, to determine commune funding, electoral boundaries, the composition of municipal councils, etc.

The "total de jure population" of a commune includes persons

- whose principal residence is within the commune,
- who live in an institution or a collective dwelling located within the commune,

- who live in an institution or a collective dwelling located in another commune but have kept a dwelling in their commune of origin,
- who live in a collective dwelling in another commune for work or live in another commune for education,
- who are attached to the commune for administrative purposes (itinerant workers, sailors and so on).

Clearly, these populations cannot be estimated until the entire territory of the commune has been covered, that is, until the detailed estimates have been produced.

3.4. Estimation of Sampling Variance

The estimates will be accompanied by a measure of their statistical quality. Work on this project began in the fall of 2001. The preferred option at this time is to use reference tables, as is done in the Canadian Labour Force Survey, for example. The sampling variances will probably be obtained by resampling the frame.

3.5. Imprecision Due to Synthesis

In the previous section, we showed how collected data will be used to produce synthetic estimates: first, an extrapolation for an "old" census, for two rotation groups (I and II, say); then directly using the census results for a third rotation group (III, say); and finally, combining extrapolations and backward projections to calibrate the last two groups (IV and V, say).

This synthesis can be formalized using a non-response model (Särndal, 1990). The annual campaign is similar to a take-all survey that has an 80% non-response rate, which is dealt with using ratio imputation. If we let s represent the whole sample, r the respondents and s-r the non-respondents, we have

$$y_{\bullet k} = \begin{cases} y_k \text{ if } k \in r \\ \hat{\beta}\, x_k \text{ if } k \in s-r \end{cases} \quad \text{with } \hat{\beta} = \frac{\overline{y}_r}{\overline{x}_r}$$

Thus, the imputation model is

$$\xi : \begin{cases} y_k = \beta\, x_k + \varepsilon_{k.} \\ E(\varepsilon_k) = 0 \\ V(\varepsilon_k) = \sigma^2 x_k \end{cases}$$

With such a model, under simple random sampling,

$$\hat{Y}_\bullet = \frac{N}{n}\sum y_{\bullet k} = \frac{N}{n}\left\{\sum_r y_k + \sum_{s-r} \hat{\beta} x_k\right\} = \ldots$$

$$= N\,\frac{\overline{y}_r}{\overline{x}_r}\,\overline{x}_s$$

The uncertainty around estimation with imputation depends on the sampling errors and the quality of imputation model ξ:

$$(\hat{Y}_\bullet - Y) \quad = \quad (\hat{Y} - Y) \quad + \quad (\hat{Y}_\bullet - \hat{Y})$$

$$\begin{matrix} \text{total} \\ \text{uncertainty} \end{matrix} \quad = \quad \begin{matrix} \text{sampling} \\ \text{uncertainty} \end{matrix} \quad + \quad \begin{matrix} \text{uncertainty} \\ \text{of model} \end{matrix}$$

This assumes that the imputation is unbiased:

$$E_\xi E_s E_r (\hat{Y}_\bullet - Y) = 0$$

Therefore,

$$V_{total} = E_\xi E_s E_r (\hat{Y}_\bullet - Y)^2 = \ldots$$
$$= E_\xi E_s E_r (\hat{Y} - Y)^2 + E_\xi E_s E_r (\hat{Y}_\bullet - \hat{Y})^2$$
$$= E_\xi V_s + E_s E_r V_\xi$$
$$V_{total} = V_{sample} + V_{imputation}$$

For many imputation models, using imputed data as if they were observed data to compute the estimate of V_s results in an underestimate of V_{sample} . In terms of expectation,

$$E_\xi \left(\hat{V}_s - \hat{V}_{\bullet s} \right) = V_{dif} \ .$$

For the estimators of these variances, Särndal shows that we get

$$\hat{V}_{sampling} = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \left\{ S_\bullet^2 + C_0 \hat{\sigma}^2 \right\}$$

with C_0 close to $\left(1 - \frac{m}{n}\right)\bar{x}_{s-r}$ and $\hat{\sigma}^2$ close to $\frac{\sum_r e_k^2}{\sum_r x_k}$ and

$$\hat{V}_{imputation} = N^2 \left(\frac{1}{m} - \frac{1}{n} \right) A \bar{x}_s \hat{\sigma}^2 \ ,$$

with $A = \frac{\bar{x}_{s-r}}{\bar{x}_r}$, which we can take as a respondent selection effect. We note that if $x_k \equiv 1$, then we do not impute, and we obtain a two-phase sample of size *m* in *n* and *n* in *N*. In addition, if *s* = *r*, $V_{total} = V_{sampling}$.

In Särndal's model, the x and y are contemporaneous; at the very least, we will have observed some of the y. Using the structure developed in the previous section, we would have:

Year *A-2*	
y_k	x_k
$y_{\bullet k}$	x_k

m respondents (Group II)
n-m imputations (other groups)

In the continuous census system, not everything is synchronous:

… A-4		A-3		A-2		A-1	A
Y_I^{A-4}	X_I^{A-4}		X_I^{A-3}		$X_I^{A—2}$		
	X_{II}^{A-4}	Y_{II}^{A-3}	X_{II}^{A-3}	$Y_{\bullet II}^{A-2}$	X_{II}^{A-2}		
	X_{III}^{A-4}		X_{III}^{A-3}	Y_{III}^{A-2}	X_{III}^{A-2}		
	X_{IV}^{A-4}		X_{IV}^{A-3}		X_{IV}^{A-2}	… …	
	X_V^{A-4}		X_V^{A-3}		X_V^{A-2}		… …

That is, Y_{II}^{A-3}, X_{II}^{A-3} , $Y_{\bullet II}^{A-2}$, and X_{II}^{A-2} are not all measured or observed in the same year. In fact, if we look at Group III on its own, for example, we have a sample of size n in year A-2 and an identical but totally non-respondent sample in year A-3. Consequently, some parameters in the estimate of V_{total} cannot be calculated.

On the other hand, if we take the problem over a specific period, we have a sample of size n and 4n non-respondents. We could approximate the uncertainty of the asynchronous imputation process (the process we have in the redesigned census) with the uncertainty of the synchronous imputation process (similar to Särndal's model).

This approach was tested on the small communes of Rhône-Alpes, for which the rotation groups, 1990 property tax data and 1990 population census data are available.

4. WORK IN PROGRESS

The methodological work involved in redesigning the census is far from complete. The following projects are still under way:

- establishment of rules for crossing the size threshold, problems of oscillation around the 10,000 population threshold, and calculation of the de jure population;
- the sensitivity of stratum boundaries in large communes and their robustness over time;
- the updating and maintenance of sampling frames and samples, especially adjustments that may be required when a commune crosses the size threshold and the incorporation of new objects into rotation groups;
- massive imputation and synthesis, both models and their precision;
- estimation of the precision of estimators; and
- collecting data from mobile population groups.

BIBLIOGRAPHY

Bertrand, P., (2000), E*stimations annuelles dans la rénovation du recensement de la population*, working paper, Département de la démographie, INSEE.

Borchsenius, L. (2000), « From a Conventional to a Register-based Census of Population », Les Recensements après 2001, Séminaire Eurostat,-INSEE, Paris.

Deville, J.C., Tillé, Y.,(1999) *Balanced Sampling by Means of the Cube Method*, CREST-ENSAI, working paper submitted for publication.

Deville, J.C., Tillé, Y. (2000), « Echantillonnage équilibré par la méthode du cube et estimation de variance » , *Journées de Méthodologie*, December 2000, INSEE, Paris.

Horvitz,D.G., (1986), « Statement to the Subcommittee on Census and Population », Committee on Post Office and Civil Service, House of Representatives, Research Triangle Park, North Carolina.

Jacod, M. and Deville J.C. (1996), « Replacing the Traditional French Census by a Large Scale Continuous Population Survey », *Annual Research Conference Proceedings*, USBC, Washington.

Kish, L., (1981), « Population Counts from Cumulated Samples », Congressional Research Service, *Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau*, Prepared for the Subcommittee on Census and Population, Committee on Post Office and Civil Service, House of Representatives, Washington.

Kish, L. (1990), « Rolling Samples and Censuses », Survey Methodology, Vol 16, N° 1, pp. 63-71, Statistics Canada, Ottawa.

Kauffmann, B., (2000), *Estimation de la précision due au modèle de synthèse*, Working paper, Département de la démographie, INSEE.

Laihonen, A. (2000), « 2001 Round Population Censuses in Europe », *Les Recensements après 2001*, Séminaire Eurostat,-INSEE, Paris.

Nathan, G., (2001), « Models for combining longitudinal data from administrative sources and panel surveys », Invited paper, ISI, Seoul, August 2001.

ONU (1990), *Principes et recommandations complémentaires concernant les recensements de la population et de l'habitat*, Etudes statistiques, ST/ESA/STA/sérieM/67, New York.

Särndal, C.E.,(1990), « Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used », *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, Ottawa, October 1990, pp. 337-347.

- (1994) « Radical Alternatives » , *Modernizing the U.S. Census*, B. Edmonston et C. Schultze, (eds.); Panel on Census Requirements in the Year 2000 and Beyond, National Research Council, National Academy Press, pp. 59-74.