Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

THEORETICAL AND PRACTICAL PROBLEMS IN CONSTRUCTING THE MSX: HOW CAN THE PRECISION OF REGIONAL EXTENSIONS OF NATIONAL SURVEYS BE IMPROVED THROUGH ADDITIONAL SAMPLING?

Marc Christine and Laurent Wilms[1]

ABSTRACT

This paper describes a method of constructing a sample for regional extensions to national surveys. It is based on the use of a balancing technique in a conditional context.

KEY WORDS: Balancing; Conditional Sampling; Inclusion Probabilities; Reverse Balancing; Regional Precision; Stratification.

1. INTRODUCTION

National household surveys in France, most of which are conducted by personal interview, are based on a sample of households selected from a master sample (MS). For the sake of simplicity, it can be said that the sample is composed of a first stage in which the primary sampling units (PSUs) are selected and a second stage in which the dwellings are selected. This master sample was designed to ensure acceptable precision for national surveys while limiting survey costs and especially travel costs by concentrating the surveys in the first stage PSUs.

Regional users regularly, and with increasing frequency in view of the successive waves of decentralization legislation in France, want regional results in order to improve precision and make comparisons between regions or between the regional level and the national level.

To meet this demand, the French national statistical institute, INSEE,[2] has decided to introduce regional extensions[3] to the national household surveys it conducts.

In the past, these survey extensions were usually administered directly by INSEE's regional directorates, separately from the national survey. As a result, the central authority deemed it necessary to construct a standard framework so that regional extension requests could be dealt with consistently and methodological issues could be managed on a global basis.

First, this framework requires that a methodology be specified for constructing the samples of households for the regional extensions. To that end, a supplementary regional sampling system was introduced. This facility, to be described below, was dubbed the master sample for regional extensions, or MSX (in French "**E**chantillon **M**aître pour les **Ex**tensions régionales", or **EMEX**). It was used for the first time in the 2002 health survey.

[1] Marc Christine, INSEE, Statistical Methods Unit F402 18, Boulevard Adolphe Pinard, F-75675 Paris cedex 14, marc.christine@insee.fr, and Laurent Wilms, INSEE, Census Renewal Program Directorate L112 18, Boulevard Adolphe Pinard, F-75675 Paris cedex 14, laurent.wilms@insee.fr .

[2] Institut national de la statistique et des études économiques.

[3] It is important to distinguish between purely local surveys, which ***are not affected by the problem discussed***, and regional extensions of national surveys, which have objectives at the regional level that are similar to the national objectives.

This paper will describe the theoretical and practical difficulties encountered in implementing the sample and the various ways of addressing them.

2. GENERAL PRINCIPLES APPLIED IN CONSTRUCTING THE MSX

As its name indicates, this tool is based on the same overall philosophy as the master sample introduced in 2000-2001. ***But the key point is MS + MSX regional representivity, which will ensure high-quality results.***

The method used to select the extensions is similar to the method used to select the dwellings from the master sample. Within strata defined by the intersection of region and level of urbanization, the sampling process consists of:

- a first stage involving the selection of specific primary sampling units (referred to as MSX PSUs), the number of which is fixed within each stratum (the same strata applied to all regional extensions);

- a second stage in which, for each survey, dwellings are selected from the MSX PSUs of extension regions[4] (usually by simple random sampling).[5]

The following general principles were applied.

- The MSX PSUs and the PSUs used for the national MS are *disjoint*, and the MSX was therefore a complement of the MS.

- In rural areas, the MSX PSUs (like the MS PSUs) are groups of communities or small towns, and in agglomerations with a population of more than 100,000, they are groups of contiguous buildings, referred to as *districts*. In small agglomerations, the agglomeration itself is a PSU.[6]

- The number of PSUs to be selected from each stratum was determined in advance by guessing the probable average size of the extensions in each region (roughly the regional fraction of the national sample).

The main problem, the solution of which is the subject of this paper, is to determine the appropriate method of selecting the MSX PSUs so as to ensure the regional representivity of the sample drawn from the MS and the MSX.

3. SELECTING THE MSX PRIMARY SAMPLING UNITS: BALANCING AND THE THEORETICAL PROBLEMS IT RAISES

3.1 Background Information on Balancing

The precision achieved by selecting samples in accordance with a particular sample design can be improved by using the technique of *balancing*.

[4] In any extension region, we have the dwelling sample selected from the MSX PSUs and the dwelling sample selected from the MS PSUs. In general, however, only a few regions have extensions (5 in the case of the 2002 health survey).

[5] In practice, the computer program that selects dwellings from the PSUs works on a global, integrated basis. It computes the dwelling allocations to be selected from each PSU (without distinguishing between MS and MSX) and performs the sampling on all eligible PSUs in each region (MS alone if there is no extension; MS + MSX if there is one).

[6] In this stratum, the MS consisted of a selection of agglomerations.

Balancing involves forcing a Horvitz-Thompson(H-T) estimator of the totals of certain variables of interest, an estimator constructed on the basis of the units selected, to take a value identical to the (known) value of the total for the entire population.

The object is to ensure that the sample selected is a better working model of the reference population for the balancing variables considered, which are assumed to be strongly correlated with the survey's variables of interest. Though the term is not strictly correct, this sampling process can be said to provide better "representivity".

We currently have balanced sampling algorithms that can be used to select from a population a sample of statistical units that meet the balancing constraints for that population, *for a predetermined set of inclusion probabilities*. One of those algorithms is the so-called CUBE method, recently devised and developed by Jean-Claude Deville and Yves Tille.

For the MS, the PSUs were selected stratum by stratum.

- *For rural PSUs*, balancing was performed at the *region group* level only (to increase the number of degrees of freedom at the time of selection), since the *PSUs' inclusion probabilities were proportional to their size*. In concrete terms, this means that a rural PSU that is in a given region and has certain socio-demographic characteristics can be represented statistically by a PSU that has similar characteristics *but is in another region in the same region group*. The variables used for balancing were age groups, taxable net income, and number of dwellings (principal or otherwise). As a result, regional representivity (and, in particular, coverage of the entire region) was not assured in advance in the MS.[7]

- *For districts in large agglomerations*, simple random sampling was used, also with balancing conditions for each agglomeration.[8]

The matter of balancing arises again in the same terms for selecting the MSX PSUs but leads to an important theoretical difficulty, which will be described below.

3.2 The Problem with the MSX: "Ex-post" Sampling

As noted above, the MS PSUs were selected in a way that made it impossible to ensure true regional representivity. This operation took place in 1999-2000. It wasn't until later that the decision was made to design and select the MSX (2001-2002). At that point, since the introduction of the MS, the organization of the surveys and the formation of the team of interviewers were halted as soon as the MS PSUs were known, *the results of the MS PSU selection process could not be questioned*.

The question that arose was as follows: ***Is it possible to select the MSX PSUs once the MS PSUs have been selected, while ensuring balancing conditions for the set MS + MSX (and not just for MSX) for each region?***

To apply balancing conditions, it is best to look at how to construct unbiased estimators of regional totals based on the union of the two samples, MS and MSX.

Clearly, with two samples for each region, one corresponding to the regional part of the MS and the other derived from the MSX, the precision of the estimate of a regional total is increased by combining *all available observations*.[9]

[7] The employment survey is an exception. It has a special (area) sample separate from the master sample, and the latter was constructed with the explicit constraint that it had to satisfy the regional precision criteria set by Eurostat.

[8] Districts are generally small and very numerous. They are fairly uniform "grains". For this reason, there was no real need to refine the selection method, and simple random sampling was used. In contrast, rural PSUs are small in number, large in size (1,800 to 3,600 principal dwellings) and generally heterogeneous.

[9] The object is to construct, from the union MS + MSX, *unbiased* estimators of totals for the region. The MS alone also achieves this aim (domain theory), though it does not ensure accurate regional representivity. However, the MSX cannot do so

The specific context of the MSX is that the new PSUs must be selected *conditionally on the particular MS selected*. Hence, a conditional sampling distribution is needed to carry out this selection process. A priori, this requires at least partial knowledge of the distribution. The distribution will be approximated (though with a loss of information) by the *first-order conditional inclusion probabilities* of the PSUs selected for the MSX.

In this context, *the sampling distribution of the MS PSUs is considered a given*; it is only partially known through the MS PSUs' first-order inclusion probabilities.

3.3 Theoretical Framework

To define a theoretical framework, let ***P*** be the set of PSUs of a given region (reference population), and let S_1 and S_2 be, respectively, the regional portion of the MS and the part of the MSX that relates to the region being considered, both of which have been selected from ***P***. Let S be the union of S_1 and S_2. For a unit i of the population, let:

Π_i^1 be the unit's probability of inclusion in sample S_1; Π_i^{2/S_1} be the unit's first-order probability of inclusion in sample S_2 *conditional on the particular S_1 selected* (it is a function of S_1 whose actual values depend on the realized s_1 of S_1); and Π_i be the unit's final probability of inclusion in the overall sample S. These probabilities are shown to be related in the following way

$$\Pi_i = \Pi_i^1 + E\,(\Pi_i^{2/S_1}\, 1_{i \notin S_1}) \tag{3.1}$$

(The expectation here is taken over the first sample's sampling distribution.)

In practice, the samples S_1 and S_2 will be *disjoint* since the MSX PSUs must be distinct from the MS PSUs. As a result, we have

$$\Pi_i^{2/S_1} = 0 \text{ for } i \in S_1 .$$

The Π_i^1 are the probabilities of inclusion in the first sample (MS); *they are given once and for all*. For rural PSUs and small agglomerations, they are proportional to size in terms of number of principal dwellings in the 1999 census of population; for districts selected from large PSUs, they are constant (equal probability sampling).

The Π_i final inclusion probabilities will also be dictated by the conditions derived from the sampling method used in later stages and by any conditions we may want to set for the weighting of the final units (dwellings), such as *final equiprobability of dwellings* (self-weighting survey) (cf. §6.2).

The parameters that we can adjust are those which define the second sample's selection conditionally on the first sample's selection, as approximated by the conditional inclusion probabilities.

3.4 Theoretical Difficulties

One difficulty lies in the fact that the above-noted relationship (3.1) between the various probabilities is complex and that for given final probabilities, it is far from simple (perhaps even impossible) to find conditional probabilities that lead to those final probabilities. In practice, that would mean knowing the first sample's entire sampling distribution and not just the first-order inclusion probabilities (except in special cases, such as the case where sample S_1 is selected by simple random sampling).

The second difficulty relates to the nature of the desired balancing constraints.

---

either, because it is drawn from a complement of the MS, which is "unbalanced" because the PSUs are selected from the MS with unequal probabilities.

In conventional balancing, the balancing condition that one hopes to attain for variable X when one selects sample S from population ***P*** is of the form $\hat{T} = T$, where T is the actual total for variable X, known for the population from which the sample is being selected, and $\hat{T}$ is the estimator of that total based on the sample.

In the context of a conventional Horvitz-Thompson estimator, the equation is as follows

$$\sum_{i \in S} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i \ .$$

The second term of this equation represents the total of variable X for the region considered; that is, *the total used in balancing must be the total of the observations for the balancing variable (perhaps a vector variable) in the population*.

It should be noted that the form of the estimator used (in this case, Horvitz-Thompson) *does not depend functionally on the balancing variables*; it is computed in such a way that the estimator obtained is an *unbiased estimator of the total of any variable Y*. This falls within the configuration where $\forall$ Y : E [$\sum_{i \in s} \frac{Y_i}{\Pi_i}$] = $\sum_{i \in P} Y_i$, with the expectation taken over the sample's sampling distribution.

For the balancing variable X, the above equation is true systematically and not just for the purposes of expectation.

In the case of interest to us, since S is the union of the disjoint samples S_1 and S_2, this condition can be written

$$\sum_{i \in S_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\Pi_i} . \tag{3.2}$$

Since sample S_1 was selected once and for all, S_2 must be selected *conditionally on the selection of* S_1, from ***P*** - S_1. Hence, the balancing conditions can be understood only within the framework of this conditional selection and with respect to a parent population, in this case, ***P*** - S_1.

The only balancing conditions that can be attained in this sampling process must therefore be written (for any balancing variable Z) in the following form

$$\sum_{i \in S_2} \frac{Z_i}{\Pi_i^{2/S_1}} = \sum_{i \in P - S_1} Z_i \ . \tag{3.3}$$

That is how the conventional framework is modified and standard tools can be used to perform this balancing process.

In the conventional framework, balanced samples can be obtained in practice by applying the algorithm from the CUBE theory using a Horvitz-Thompson estimator.

When modified to the present case, in the context of conditional sampling, this method, in order to be used, would require the following relationship to be true

$$\mathrm{E}\,[\sum_{i \in S_2} \frac{Y_i}{\Pi_i^{2/S_1}} \ / \ \mathrm{S}_1] = \sum_{i \in P - S_1} Y_i \ ,$$

for any variable Y whose total is estimated on the basis of the second sample's conditional selection (expectation here is taken *over the conditional distribution of S_2 given S_1*), where sample S_2 *is selected in such a way that there is actual (not just expected) equality for the balancing variable Z.*

But since the Π_i, unit i's probabilities of inclusion in the overall sample S, are predetermined, it is generally impossible to find Π_i^{2/S_1} that will satisfy equations (3.2) and (3.3) and relationship (3.1) above.

Clearly, except in very special cases, which will be discussed in §4, the problem in question does not meet the conditions required in order to apply CUBE sampling in a conventional balancing process.

Thus, to find a general solution to this problem, it will be necessary to resort to approximate solutions.

4. SPECIAL SITUATIONS IN WHICH THE PROBLEM HAS AN EXACT SOLUTION

There are three special cases in which the problem can actually be solved.

4.1 Simultaneous Selection of the MS and the MSX

If the MSX problem had been dealt with at the time the MS was selected, and the two samples had been selected simultaneously, there would have been a very simple way of balancing them simultaneously on a variable X, using the procedure described below.

(a) Select a sample S of k units from ***P*** with inclusion probabilities Π_i, while balancing on X

$$\sum_{i\in S}\frac{x_i}{\Pi_i} = \sum_{i\in P} x_i \ .$$

This sample S corresponds to the union of the MS and the MSX.

(b) Select from S, by simple random sampling at rate f, a sample S_2 of fixed size (k - a), which will represent the MSX, while balancing on the variable $Z_i = \frac{x_i}{\Pi_i}$

$$\sum_{i\in S_2}\frac{1}{f}\frac{x_i}{\Pi_i} = \sum_{i\in S}\frac{x_i}{\Pi_i} \ .$$

(c) Then the sample S - S_2, which would represent the MS alone, gives each unit i the selection probability Π_i^1 = (1 - f) Π_i. In particular, if Π_i is proportional to the size of unit i, the same is true for Π_i^1. Moreover, it is easily verified that

$$\sum_{i\in S-S_2}\frac{x_i}{(1-f)\Pi_i} = \sum_{i\in P} x_i \ ,$$

i.e., that S - S_2 is also balanced on the same variable X for population ***P***.

In concrete terms, this means that a larger sample, balanced on the variable of interest, is selected systematically from the outset; in regions with a regional extension, it is this sample that is used (representing MS + MSX), while in other regions, the useful sample (the MS) is selected in a second phase, by simple random sampling.

The disadvantage of this procedure is twofold:

- First, it could only have been applied by deciding on the selection of both samples in advance (whereas in reality the necessity and value of the MSX did not emerge until *after* the MS was selected).

- Second, for regions with no extension, selecting the MS in two phases instead of directly could affect the variance.

4.2 Cases of Successive Balanced Sampling with Equal Probabilities

As noted above (§3.1), this case was encountered in the selection of districts in large agglomerations.

This case is subject to the theoretical result below (proof of which is left to the reader).

Proposition:

If a sample S_1 is selected from a population ***P*** according to a sampling process balanced on a variable X, and if the sampling is done with *equal probabilities*; and If a second sample (S_2) is selected from ***P*** - S_1 such that the conditional selection of S_2 given S_1 is with equal probabilities and balanced on the same variable X (for ***P*** - S_1). Then $S = S_1 \cup S_2$ is balanced on X for the total population ***P*** (with respect to the final sampling distribution of S).

4.3 Using a Special Form of the Horvitz-Thompson Estimator

Take $\hat{T} = \sum_{i \in S_1} Y_i + \sum_{i \in S_2} \frac{Y_i}{\Pi_i^{2/S_1}}$ as an estimator of the total of any variable Y. This produces an unbiased estimator. In addition, when S_2 is selected from ***P*** - S_1 by a process that is balanced on X, the estimator extends the property of being balanced on X to the population ***P***, for the overall sample S.

It is as if S_1 were a reference population: one takes the unweighted total for that sub-population and adds a Horvitz-Thompson estimate using the second sample's conditional sampling distribution. This ignores the inferential properties of sample S_1, which is certainly not ideal.

5. BACK TO THE GENERAL CASE: AN APPROXIMATE SOLUTION, *REVERSE BALANCING*

Let us assume that we are using a Horvitz-Thompson estimator on the union of two samples, with *fixed* final inclusion probabilities Π_i.

The equation for balancing on variable X should be:

$$\sum_{i \in S_1} \frac{x_i}{\Pi_i} + \sum_{i \in S_2} \frac{x_i}{\Pi_i} = \sum_{i \in P} x_i \ . \tag{5.1}$$

Since we cannot be certain of this equation in advance for a given distribution of the first sample, we will relax one constraint: we will allow ourselves to deviate "to an acceptable extent" from the desired value of the final inclusion probabilities Π_i, instead of keeping it fixed. In other words, we will be looking for new final inclusion probabilities, $\tilde{\Pi}_i$, that are close to the Π_i, so that, by introducing an ad hoc conditional sampling distribution, we can satisfy the general balancing equation (5.1). In concrete terms, we will attempt to minimize a parameter of the form $\sum_{i \in P} d(\tilde{\Pi}_i, \Pi_i)$, while keeping to balancing equation (5.1), where ***P*** is the set of PSUs forming a stratum (intersection of a region and an agglomeration category) and *d* denotes a distance (Euclidian distance, or χ^2 distance in practice). The balancing equation must be valid in the context of the conditional selection of S_2 given S_1 (since the MS PSUs have already been selected). What is needed, then, is a form equivalent to equation (5.1), which can actually be interpreted as a balancing condition to be met in the conditional selection of S_2 given S_1. To that end, we will write equation (5.1) in the form:

$$\sum_{i \in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i}$$

where we let $z_i = x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i}$ and $\tilde{\Pi}_i^{2/S_1}$ denotes a new conditional sampling probability for PSU i given S_1.

For this equation to be interpreted as a true balancing constraint on the total of the z_i variables for the population within which the conditional sampling is being done (***P*** - S_1), it must be written in the form

$$\sum_{i \in S_2} \frac{z_i}{\tilde{\Pi}_i^{2/S_1}} = \sum_{i \in P - S_1} z_i \ . \tag{5.2}$$

These conditions are satisfied if and only if

$$\sum_{i \in P} x_i - \sum_{i \in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i \in P - S_1} z_i = \sum_{i \in P - S_1} x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i} \ . \tag{5.3}$$

Let us take a <u>simple case of determining the conditional sampling distribution</u> of S_2 given S_1, defined as follows:

$$\Pi_i^{2/S_1} = \begin{cases} 0 & \text{if } i \text{ is a member of } S_1 \\ \mu_i & \text{if } i \text{ is not a member of } S_1 \end{cases}$$

that is, Π_i^{2/S_1} depends on S_1 only through the indicator $\boldsymbol{1}_{i \in S_1}$. The μ_i coefficients depend on unit i, but their value is fixed in advance, independent of the particular instance of sample S_1.

On the basis of equation (3.1), this yields

$$\begin{aligned} \Pi_i &= \Pi_i^1 + \mu_i \sum_{S_i / i \notin S_1} P(S_1 = s_1) \\ &= \Pi_i^1 + \mu_i (1 - \Pi_i^1) \end{aligned}$$

or

$$\Pi_i = (1 - \mu_i)\ \Pi_i^1 + \mu_i$$

Rearranging, we get

$$\mu_i = \frac{\Pi_i - \Pi_i^1}{1 - \Pi_i^1} \ . \tag{5.4}$$

Since Π_i^1 is fixed, there is, in the determination of the conditional distribution, a one-to-one correspondence between the Π_i and the Π_i^{2/S_1}.

For example, if we want the PSUs' final probabilities of inclusion in the union of the two samples to be, as they are for the MS alone, proportional to their size, with fixed sample size k (within a given stratum), the following condition would apply

$$\forall\ i \in \boldsymbol{P} : \Pi_i = k \frac{T_i}{T}$$

where T_i is the size of unit i and T is the size of the total population (the sum of the unit sizes).

With an analogous condition for the MS – $\Pi_i^1 = a \frac{T_i}{T}$, where a is the number of units selected from the MS – we obtain the relationship

$$\mu_i = \frac{(k-a)T_i}{T-aT_i}.$$

This generally yields acceptable probabilities, with values in [0,1].[10]

When the actual final inclusion probabilities Π_i are replaced with the approximate probabilities $\tilde{\Pi}_i$, the conditional probabilities that produce the $\tilde{\Pi}_i$ must be adjusted uniformly while maintaining relationship (5.4).

This leads to the following relationship, for all i that are not members of S_1

$$\tilde{\Pi}_i^{2/S_1} = \frac{\tilde{\Pi}_i - \Pi_i^1}{1-\Pi_i^1}.$$

Hence, constraint equation (5.3) can be written solely as a function of $\tilde{\Pi}_i$ as

$$\sum_{i\in P} x_i - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i\in P-S_1} x_i \frac{1-\frac{\Pi_i^1}{\tilde{\Pi}_i}}{1-\Pi_i^1}.$$

Finally, in this case, the problem comes down to solving the $\tilde{\Pi}_i$ search program (P):

$$\text{Min} \sum_{i\in P} d(\tilde{\Pi}_i, \Pi_i)$$

subject to the constraint that

$$\sum_{i\in P} x_i - \sum_{i\in S_1} \frac{x_i}{\tilde{\Pi}_i} = \sum_{i\in P-S_1} x_i \frac{1-\frac{\Pi_i^1}{\tilde{\Pi}_i}}{1-\Pi_i^1}.$$

It may also be necessary to add the fixed-size constraint given by $\sum_{i\in P} \tilde{\Pi}_i$ = k. In the end, the solution of this program leads to new final inclusion probabilities $\tilde{\Pi}_i$, from which we will deduce conditional inclusion probabilities $\tilde{\Pi}_i^{2/S_1}$ that satisfy the properties described below.

If we select the second sample on the basis of these inclusion probabilities while balancing it on the total of the variable $z_i = x_i \frac{\tilde{\Pi}_i^{2/S_1}}{\tilde{\Pi}_i}$ for population ***P*** - S_1, then, when we combine the two samples, we not only give each unit i an overall inclusion probability equal to $\tilde{\Pi}_i$, but also perform balancing on the total of the variable X for the entire population ***P***.

An important difference to note relative to the conventional balancing procedure is that in the latter, any sample can be selected with inclusion probabilities ***fixed in advance*** by requiring the sample to satisfy balancing constraints. The procedure described here does the "reverse": it uses inclusion probabilities that will satisfy the balancing equation and ***will therefore depend on the balancing variables selected*** and the values they take in the population. ***This is why the procedure is referred to as "reverse balancing".***

[10] Except in certain special cases where some T_i are very large.

Another significant difference relates to *calibration procedures*. Those procedures involve adjusting the initial weights (equal to the inverse inclusion probabilities) while attempting to minimize a particular distance between initial and final weights, in order to force the estimators to satisfy calibration constraints. However, the new weights are computed ex-post *on the selected units only*. In reverse balancing, on the other hand, final inclusion probabilities are computed for all units in the population.

Nevertheless, reverse balancing raises the question of whether there are acceptable solutions (with all $\tilde{\Pi}_i$ in $[\Pi_i^1, 1]$); the solution is obtained by numerical methods.[11]

6. CONCLUSION: THE VALUE OF THE MSX AND PROSPECTS FOR IMPROVEMENT

This study leads to two types of conclusions.

6.1 Practical Conclusions

The MSX combines extension samples with the national sample so that only one weight calculation is necessary and the resulting weights can be used in preparing both national and regional estimates and in standardizing processing methods. However, since the MSX was not introduced until 2002, it would be premature to attempt a full assessment of it at this stage.

6.2 Theoretical Conclusions

The paper sheds light on the difficulty of obtaining a balancing property in a successive sampling environment and clarifies the notion of conditional sampling. It offers an acceptable approximate solution to the problem.

6.3 The Future: Simultaneous Design

As we saw, the problem could have been solved very easily by selecting the MS PSUs and the MSX PSUs simultaneously (cf. §4.1), and the difficulty arose from the fact that the need for regional extensions did not emerge until *after* the national master sample had been selected.

In the context of INSEE's current research into constructing new samples based on the redesigned census of population (which uses a rotational sampling system), this concern with being able to respond to regional extension requests by establishing a complementary sampling process will be taken into account in advance.

As a result, the situation in the future will be different, and the theoretical problems described in this paper should not arise.

REFERENCES

Bourdallé G., Christine M., Wilms L., « Echantillons Maître et Emploi », *Insee-Méthodes : VII èmes JMS, 4-5 décembre 2000*, n° 100 (tome 1), p. 139-241.

Bousabaa A., Lieber J. et Sirolli. R (1999) : la macro Cube, Document de travail, ENSAI, Rennes.

Deville J-C. et Tillé Y. (2001) : variance reduction using balanced sampling : the Cube method.

Favre A-C.et Tillé Y. : coordination, combination and extension of optimal balanced samples, à paraître.

[11] Using an SAS/IML procedure.

Rousseau S. et Tardieu F. (2004) : la Macro SAS CUBE d'échantillonnage équilibré, documentation de l'utilisateur. *Cette documentation peut être téléchargée sur le site de l'Insee : www.insee.fr, rubrique « nomenclatures, définitions, méthodes ».*