Survey Methodology, December 1997
Vol. 23, No. 2, pp. 91-98
Statistics Canada

A Synthetic, Robust and Efficient Method of Making Small Area Population Estimates in France

GEORGES DECAUDIN and JEAN-CLAUDE LABAT[1]

ABSTRACT

Since France has no population registers, population censuses are the basis for its socio-demographic information system. However, between two censuses, some data must be updated, in particular at a high level of geographic detail, especially since censuses are tending, for various reasons, to be less frequent. In 1993, the Institut National de la Statistique et des Études Économiques (INSEE) set up a team whose objective was to propose a system to substantially improve the existing mechanism for making small area population estimates. Its task was twofold: to prepare an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is reported on here, is flexible and reliable, without being overly complex.

KEY WORDS: Population estimates; Administrative files; Robust estimation.

1. INTRODUCTION

In France, as in all countries that do not have population registers, censuses of the population are the cornerstone of the socio-demographic information system. However, censuses are quite massive operations that cannot at present be carried out more often than once every seven or eight years. In the interval between censuses, it is therefore necessary to update some information, especially at a high level of geographic detail, particularly since for various reasons, censuses are tending to be less frequent. Thus, small area population estimates are a major challenge for the Institut National de la Statistique et des Études Économiques (INSEE).

Despite the progress achieved in this field, the situation in 1993 still seemed fairly unsatisfactory. When figures from the 1990 population census were compared to the population estimates made on the basis of the previous census (1982) for the metropolitan departments, the differences noted were sometimes sizable.

INSEE therefore created a methodology team whose mission was to propose a system that would substantially improve the existing mechanism. Initially, the next census was to take place in 1997. It therefore seemed reasonable to have the new system operate on an experimental basis until the census, so as to see how well it worked before using it in actual production. When the census was postponed to 1999, it became more necessary to bring the project to a successful conclusion quickly, so as to be able to use the new system in 1996.

To achieve its objective, the team devoted itself, with maximum pragmatism, to a twofold task: to develop an efficient and robust synthesis of the information available from different administrative sources, and to assemble a sufficient number of "good" sources. The "multi-source" system that it designed, which is described here, is not overly complex and seems effective. A more detailed description of it is provided in Decaudin and Labat (1996).

2. MAIN CONCLUSIONS

The team's main conclusions are as follows:

1) It is impossible to improve total population estimates using sample surveys, unless the survey is conducted on such a scale that it would be similar to a census.
2) No single administrative source adequately reflects changes in the population. At the local level, all sources can exhibit drift, breaks, jolts, *etc.*, which are not always easy to detect. Furthermore, even at the local level, it is often quite difficult if not impossible to get the agency responsible to provide explanatory details, much less corrections in the case of errors. In any event, it is unwise to rely on a single administrative source, however good it may be, since its permanency is never guaranteed.
3) On the other hand, total population estimates can be improved substantially by simultaneously using several sources. A "multi-source" system, similar to the one presented here but more rudimentary, was tested retrospectively over the intercensal period 1982-1990, for the 96 metropolitan departments. The mean error (mean deviation as an absolute value from the results of the March 1990 census) fell below 0.9%, whereas the mean error registered at the time, with the estimation system then in place, was 1.4%.

3. SIMULTANEOUS USE OF SEVERAL SOURCES

For using several sources jointly, different methods are possible.

A method that is universal – and easy to implement – is multiple regression. In simplified form, this amounts to using, for any area z, the following relationship:

$$P(n+1,z)/P(n,z) = c + \sum_{S} (k_S N_S(n+1,z)/N_S(n,z)),$$

[1] Georges Decaudin and Jean-Claude Labat, Institut National de la Statistique et des Études Économique, 18, Blvd. Adolphe-Pinard, 75765 Paris, CEDEX 14.

where $P(n,z)$ is the population of area z on January 1 of year n, the values $N_S(n,z)$ are the numbers from each source S on the same date and k_S are coefficients, which are estimated by multiple regression over a past period. Here c is a constant term that is used only in the regression, with calibration on the national population serving to correct any drift.

This method is used in various countries, including Canada and the United States (for example, see Statistics Canada 1987 and Long 1993). Nevertheless, it was not adopted because it has numerous drawbacks:

- it must be possible to estimate the coefficients, which requires data from each source extending back over a fairly long period;
- the coefficients can change over time, without it being possible to control this change;
- as noted above, the administrative sources are, for various reasons (changes in regulations, abrupt shifts in management, errors, *etc.*), subject to what might be called "anomalies". For each source S, the scope of these anomalies is reflected in part in the coefficient k_S, to an extent that depends on how great their medium-term effect has been over the calibration period [la période d'étalonnage]; but anomalies nevertheless occur in estimates with the same weight as the "good" data from the same source. The estimates are then highly distorted.

Another method is known as the "*composite*" method. Each source is used to estimate the population in one or more age classes: age class X, which is well-covered by the source, but also sometimes another class that definitely exhibits a pattern very similar to that of class X (for example, the "30-45" age group, if X represents the "under 18" age group). It is then necessary to have appropriate indicators for the other components of the population and correctly manage the consolidation of these estimates "in parts".

This type of method, used in the United States (Long 1993), seemed to us to be problematic, especially because of the difficulty of adequately dealing with "anomalies".

The *proposed "multi-source" system* is based on a robust synthesis of estimates from different sources. It combines demographic reasoning with purely statistical techniques. It draws on the experiments conducted by the INSEE's regional directorate in Brittany in the early 1970s (Laurent and Guéguen 1971; Guéguen 1972). Should one of the sources fail, such a system is not prevented from functioning, even though its performance may be somewhat diminished.

4. DEMOGRAPHIC BASE

The demographic reasoning which is at the base of the system is elementary: assuming that we know the total population $P(n)$ for an area on January 1 of year n, the population $P(n+1)$ of the area on January 1 of year $n+1$ is deduced by summing the two components of the change during year n: natural increase (births minus deaths), and net migration (immigrants minus emigrants).

$$P(n+1) = P(n) + N(n) - D(n) + I(n) - E(n).$$

In France, natural increase data are provided annually at the commune level by vital statistics. If the latter are not yet available in final form, which is often the case in the third quarter of year $n+1$, it is easy to estimate them with a low margin of uncertainty.

The only unknown, then, is net migration for year n: $SM(n) = I(n) - E(n)$ or what amounts to the same thing, the net migration rate $T(n) = SM(n)/P(n)$. In other words, estimating the population comes down to estimating net migration since the last date on which the population is known (or is assumed to be known), and vice versa.

In France, net migration figures are of some importance, although less so than in other countries such as Canada or the United States. In addition, they generally exhibit a certain inertia, at least at relatively aggregated geographic levels. One way to assess the influence of changes to them from one intercensal period to the next is to measure the errors that would have been committed during each period if the population had been estimated by using the average annual net migration rates for the preceding period. Over the period 1982-1990, for the departments (excluding Corsica), the mean end-of-period error (in 1990, at the end of eight years) would have been only 1.3%. It was not certain, when the team started its work, that much greater accuracy could be achieved. However, both in 1975 and in 1982, the mean error that would have been committed with the trend method would have been much greater: 2.8% and 2.7% respectively (over seven years). It would therefore seem that the period 1982-1990 was exceptional and that in the future the difference will again be more pronounced.

5. ESTIMATES FROM THE DIFFERENT SOURCES

From each source, using an appropriate method, we draw an estimate of annual net migration rate for the population as a whole. The methods that may be used depend on the data available.

For each of the sources tested and found to be "good", at least at the departmental level, a method is proposed. The five sources retained are the following: housing tax; electrical utility customers; children receiving family allowances; educational statistics; electoral file.

The data on the composition of households for tax purposes, which appear in the income tax files, are the sixth source that should provide very good results. However, to date, these data have been analysed for only a few departments, and the methodology for using them is not yet completely defined.

We also propose to integrate a trend estimate of the net migration rate into the system.

Two categories of methods are used. The first concerns the sources relating to households; the second concerns those relating to individuals.

5.1 Sources Relating to Households

Some sources provide information on changes in the number of households. This is the case with the files on *housing taxes* (HT) and *electrical utility customers* (EUC). The housing tax is one of the four main local direct taxes. As its name indicates, it applies to occupied dwellings, with main residences and secondary residences being treated separately. The housing tax file takes account of the situation on January 1 of the taxation year. Starting in the 1980s, the HT source was the basis for the departmental population estimates developed by INSEE (Descours 1992). In the early 1990s, it was replaced by the EUC source, in light of the distortions caused by a change to the HT management system which gradually worked its way through all departments.

The method adopted for using these sources follows classical principles. It leads directly to an estimate of the total population, and it involves three main stages:

1) estimating the number of households;
2) estimating average household size and from there, estimating the population of households;
3) adding the "non-household" population.

In the first stage, it is assumed that the number of households changes in accordance with the data supplied by the source (number of main residences for HT purposes or number of electrical utility customers). The second stage is more delicate. It is based on both the use of statistics on dependants from the HT files and on a trend estimate of average household size.

In the proposed "multi-source" system, we move on to the net migration rate, for comparison with other sources, using vital statistics data (*cf.* Section 4).

5.2 Sources Relating to Individuals

The other sources used concern individuals. Only a certain age group X of the population is generally covered adequately. The method then involves two main stages:

1) estimating, from the source, the net migration rate for the population aged X;
2) from there, estimating the net migration rate for the population as a whole.

The second stage is based on the following statistical relationship, observed in the past, between the change, from one period to another, of the overall net migration rate (T) and the change in the net migration rate for the population aged X (TX):

$$T_2 - T_1 = \delta_X (TX_2 - TX_1),$$

where δ_X is a coefficient close to 1, depending on the age group X. This relationship is similar to the one used by de Guibert-Lantoine (1987) to estimate the population on the basis of educational statistics.

For the corresponding age groups in the different sources used, the values, estimated by linear regression, of the coefficient δ_X(+/−2 standard deviations) are shown in tables 1 and 2.

Table 1
Estimates of δ_X on Departments, Excluding Corsica, Internal Net Migration

Period 1	Period 2	Age at end of period		
		0-19	10-14	35 and over
1962-1968	1968-1975	0.76 (+/- 0.04)	0.69 (+/-0.06)	1.24 (+/-0.09)
1968-1975	1975-1982	0.77 (+/-0.03)	0.88 (+/-0.06)	1.56 (+/-0.08)
1975-1982	1982-1990	0.70 (+/-0.11)	0.49 (+/-0.10)	1.26 (+/-0.17)

Table 2
Estimates of δ_X Over the Two Periods 1975-1982 and 1982-1990, Excluding Corsica, Total Net Migration

	Age at end of period		
	0-18	9-15	35 and over
Departments	0.65 (+/-0.11)	0.57 (+/-0.10)	1.22 (+/-0.16)
Department – employment zone	0.65 (+/-0.04)	0.59 (+/-0.04)	1.17 (+/-0.06)

The approach followed in the first stage depends on the source:

Electoral File

Annual migration figures for voters in the selected age group (30 and over) are supplied directly by the electoral file managed by INSEE. We go from the rate of net migration of voters to the residential net migration rate by dividing the former by a coefficient reflecting the magnitude of the change in the electoral file.

Educational Statistics

The net migration figure for those in the 5-9 age group is obtained by subtracting their number in year n from that of the same cohorts the next year (that is, from those in the 6-10 age group in year $n + 1$) and deducting deaths.

Children Receiving Family Allowances

The number of persons in the 0-17 age group is estimated on the assumption that it evolves similarly to the number of children receiving family allowances. From this a figure for the net migration of young persons is obtained by comparing this estimate to a hypothetical change in the youth population without migration, that is, a change due solely to natural increase.

6. SYNTHESIS

6.1 Principles

The different basic estimates of the annual net migration rate are treated statistically in order to obtain a "synthetic rate", to be used as the final estimate. The treatment serves to eliminate outliers, underweight suspect values and, more generally, assign to each source a weight that reflects its performance.

More specifically, since each source can "drift", the different basic estimates are generally biased; they are first corrected for the national bias of the corresponding source for the year considered, a bias that is estimated in advance. In proceeding in this way, we implicitly assume that the difference between the local bias and the national bias is minor in relation to the irreducible unexplained portion of the difference (flou irréductible). Once we have estimates for a number of years, it should be possible to test this hypothesis and if necessary, replace it with one that corresponds more closely to reality, so as to improve the correction of biases at the local level.

It should be noted that such a seemingly simple operation as correcting the national bias nevertheless requires several precautions. The solution that consists in carrying out a gross calibration on the national net migration rate, considered by definition as a good reference, is not very satisfactory, owing to anomalies that may distort the calibration. It is therefore preferable to estimate the biases by means of a process in which we also eliminate anomalies. The process is similar to the one used for synthesis, which is described below. However, the determination of biases, assumed to be national in scope and therefore calculated for 96 departments, is less sensitive to anomalies than the determination of synthetic rates, calculated over a small number of sources. Only major anomalies are likely to significantly throw off the calibration of the rates and must therefore be corrected.

The "synthetic" net migration rate is a weighted mean of the basic estimates thus calibrated. Each source S is assigned an initial weight W_S that is supposed to reflect its medium-term accuracy. But in addition, for a given year and area, this weight is modulated to take account of the plausibility of the corresponding rate. Thus, if a rate is "abnormally distant" from the rates obtained from other sources – in practice, from a central value for all rates for the area – its weight is cancelled or reduced. For this, we look at the distance between the rate obtained from each source and the central value identified, and we compare it to a "norm" of distance NO_S specific to the source, determined empirically on the basis of the data available: if the distance is less than "a times the norm", the weight is not automatically changed; if it is greater than "b times the norm", it is set at 0; between the two, the weight is multiplied by a coefficient, included between 0 and 1, calculated by interpolation.

Note that the trend estimate is formally treated like those from exogenous sources; its weight is cancelled when it is considered as implausible because it is too far from the other estimates.

The synthesis is achieved automatically, which ensures homogeneity and an explicit logic to the treatments carried out. This does not, however, eliminate the need to control the results obtained.

6.2 Theoretical Presentation

On the theoretical level, we sought to use reasonings and robust estimation techniques, such as described in Hoaglin, Mosteller and Tukey (1983). The method adopted falls within the framework of M-estimators of central tendency and more specifically in the category of W-estimators, which use the reweighted least squares algorithm.

Since the net migration rates for year n and area z obtained from different sources S (and corrected for their national biases) are denoted $TC_S(n, z)$, the synthetic rate $T(n, z)$ solves the implicit equation:

$$\sum_S W_S \cdot NO_S \cdot \Psi(\frac{TC_s(n,z) - T(n,z)}{NO_S}) = 0,$$

where the function Ψ is of the type that redescends to a finite rejection point:

$$\Psi(r) = r \qquad \text{for } |r| \le a,$$

$$\Psi(r) = r\frac{b - |r|}{b - a} \qquad \text{for } a < |r| \le b,$$

$$\Psi(r) = 0 \qquad \text{otherwise.}$$

Using an iterative process, we can gradually refine the automatic processing of suspect data.

6.3 First Analysis of the Distances From Each Rate to the Central Value for the Rates

1) For each area z we calculate a first central value of the "calibrated" rates $TC_S(n, z)$. The central value used must not be overly sensitive to the possible existence of quite distant values for some sources, but at the same time it must be influenced by a source to the extent that the source is on average more accurate. Under these conditions, rather than choosing the median – which would meet the first condition – we use a statistic of rank that is a little more elaborate but nevertheless simple, owing to the small number of values; this statistic is the mean, weighted by respectively 1/2, 1/4, 1/4, of the three quartiles:
   - the median of the rates $TC_S(n, z)$ weighted by the initial weights W_S,
   - the lower quartile (Q1) of the weighted rates,
   - the upper quartile (Q3) of the weighted rates.

2) The rates $T1(n, z)$ thus obtained are calibrated on the net migration rate for the higher level, by simple translation:

$$TC1(n,z) = T1(n,z) + TREF(n) - \sum_z (T1(n,z)P(n,z)) \Big/ \sum_z P(n,z)$$

where $P(n, z)$ is the population of area z on January 1 of year n and $TREF(n)$ is the net migration rate for the higher level (the national rate for the departmental synthesis).

3) For each area, we calculate the differences between each rate and this calibrated central value:

$$EC1_S(n, z) = \mid TC_S(n, z) - TC1(n, z) \mid.$$

4) For each source and each area, the size of this difference is assessed in relation to the "norm" of distance NO_S specific to the source. This "norm" is determined empirically on the basis of the available data: theoretically it is the average of the distances observed in the past, excluding anomalies. The result is a first modulation of the weight originally assigned to this source:
   - if $EC1_S(n, z) \leq a1\, NO_S$, where $a1$ is a parameter to be chosen (in the vicinity of 2), we do not change W_S, the initial weight for S. In other words, if $WM1_S(n, z)$ is the modulation coefficient of W_S (coefficient included between 0 and 1), we take $WM1_S(n, z) = 1$;
   - if $EC1_S(n, z) > b1\, NO_S$, where $b1$ is another parameter (in the vicinity of 3), we set W_S at 0, meaning that we eliminate source S: $WM1_S(n, z) = 0$;
   - if $a1\, NO_S < EC1_S(n, z) \leq b1\, NO_S$, we interpolate $WM1_S(n, z)$ as a function of the value of $EC1_S(n, z)$:

$$WM1_S(n, z) = (b1\, NO_S - EC1_S(n, z))/((b1 - a1)NO_S).$$

5) At the end of this first phase, we therefore have new weights specific to each source and each area, which would allow us to locally eliminate or underweight suspect rates: $W1_S(n, z) = W_S WM1_S(n, z)$.

6.4 Iterations

1) Using the weights thus modified $W1_S(n, z)$, we estimate a new central value for each area, this time taking the weighted average of the rates:

$$T2(n, z) = \sum_S (TC_S(n, z) W1_S(n, z)) \Big/ \sum_S W1_S(n, z).$$

2) We calibrate each rate $T2(n, z)$ on the net migration rate for the higher level, by translation. We obtain $TC2(n, z)$.
3) We calculate, in each area, the differences between each rate and the calibrated average rate: $EC2_S(n, z) = \mid TC_S(n, z) - TC2(n, z) \mid$. Using these differences, we calculate new modulation coefficients for the initial weights, using the parameters $a2$ and $b2$, which may be different from $a1$ and $b1$ (theoretically they would be lower). We thus obtain new weights $W2_S(n, z)$ which more effectively take account of anomalies, since the latter are assessed in relation to a better central tendency. With these weights, we estimate a new synthetic rate $T3(n, z)$, which is calibrated on the higher level to obtain $TC3(n, z)$.
4) The operations described in point 3 are repeated with the same parameters $a2$ and $b2$. The tests conducted at the departmental level over the period 1982-1990 show that the convergence is generally rapid; the rates are quite often stabilized by the fourth iteration.

7. IMPLEMENTATION AT THE DEPARTMENTAL LEVEL

The estimation system outlined above, which is operationalized for 1990 and subsequent years, was implemented by the project team for the year 1990 at the departmental level, with the following five sources: housing tax (HT), electrical utility customers (EUC), family allowances (FA), educational statistics (ES), electoral file (EF), plus the trend estimate (TREND).

Figure 1 shows the results obtained for several departments. Table 3 shows the values of the weights and norms used to make the system operate. This table also shows certain statistics obtained from the synthesis of the net migration rates; in particular they concern the differences between the rates obtained from each source and the synthetic rates.

Table 3
Implementation for Year 1990 at Department Level
Parameters and Statistics

	HT	EUC	FA	ES	EF	TEND
Weight	115	100	80	70	80	100
Norm	0.15	0.17	0.19	0.20	0.19	0.12
Number of rates	96	96	89	96	94	96
Average distance	0.55	0.14	0.30	0.19	0.14	0.13
Number of "aberrant" rates	37	2	17	3	1	6
Average of distances without "aberrant" rates	0.15	0.13	0.16	0.16	0.13	0.11

Note: - Coefficients (a; b) applied to norms: (2,5; 3,5) in the first iteration, then (2; 3).
- The values of the distances and norms correspond to rates expressed as a %.
- Distances are calculated in relation to the synthetic rates after three iterations.
- "Aberrant" rates are those for which the weight is cancelled after three iterations.

The results suggest that the system is even more effective than indicated by the summary retrospective test carried out on the 1982-1990 intercensal period with the same sources. Aside from the HT source, which is still distorted, the estimates from the different sources are more convergent than they were on average in the retrospective test (see Table 4).

There is nothing surprising about this, given the rudimentary state of the system tested on the 1982-1990 intercensal period. The data used were rough or even

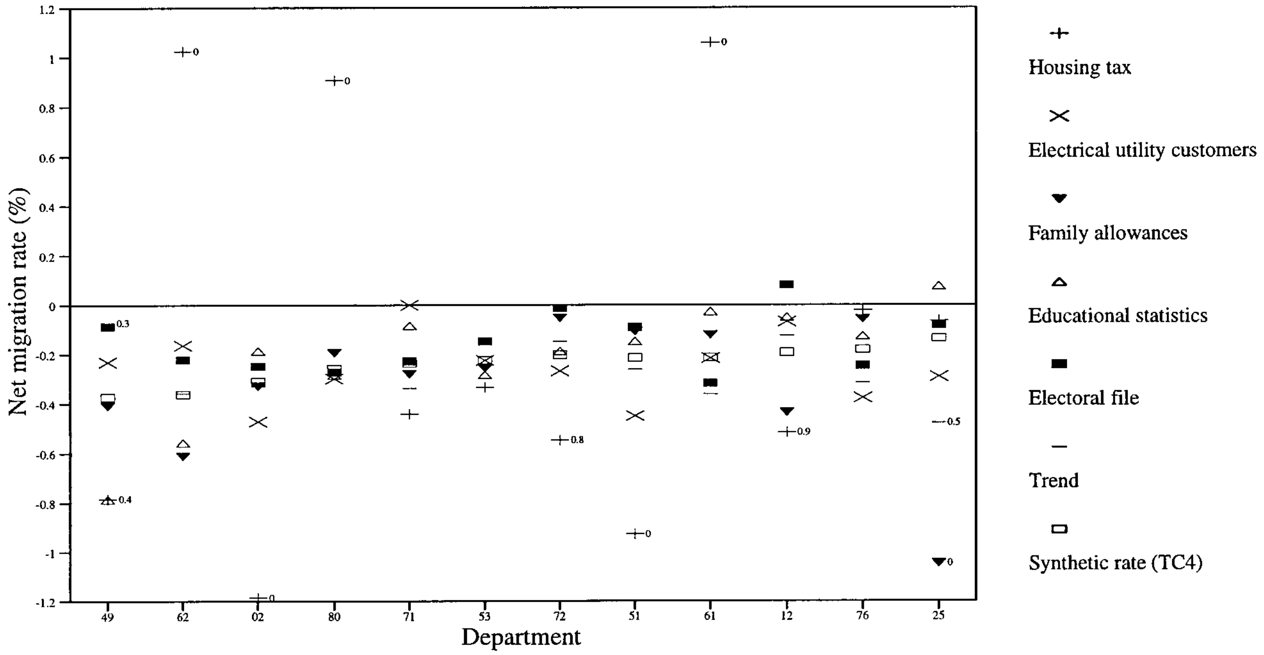

Figure 1: Summary of Net Migration Rates for 1990 for Twelve Departments, Identified by Number (49, 62, *etc.*).
Note: TC4 is the synthetic rate obtained after three iterations. Where the weight for a source has been eliminated or reduced, the value of the modulation coefficient (WM3) is shown.

fragmentary, owing to the difficulty of assembling, in 1993, management data for years past (1982, ...); in addition, the relationships used to draw an estimate of the net migration rate from each source were simplistic; and lastly, the method of synthesis was less elaborate.

It should be noted that the integration of other sources – income tax data in particular – can only further reinforce the effectiveness of the system.

Table 4
Mean of Distance in Retrospective Test

	TH	EDF	AF	EN	FE
1982	0.26	0.34	0.50	0.47	0.34
1983	0.28	0.33	0.48	0.47	0.32
1984	0.23	0.28	0.40	0.45	0.34
1985	0.24	0.31	0.48	0.44	0.32
1986	0.23	0.33	0.40	0.33	
1987	0.40	0.28	0.41	0.27	
1988	0.84	0.29	0.30	0.37	0.24
1989	0.97	0.21	0.30	0.33	0.35
Overall mean	0.43	0.30	0.41	0.39	0.32

Notes: -The number of rates per year is generally 96, except for FA (89) and EF (94).
-The "electoral file" source did not provide rates for 1986 or 1987.
-The "housing tax" source began to be distorted in 1987.
-The values of the differences correspond to rates expressed as a %.

8. SUPPLEMENTS

8.1 Sub-Departmental Levels

The use of some sources may become risky at a geographic level below the departmental level. There are various reasons for this: because the hypotheses on which the method is based become fragile, because the numbers are small, *etc.* This is especially the case with educational statistics.

However, it should be possible to operate the system for employment areas, or more specifically for cross-tabulations of department and employment area (there are approximately 420 such areas), which serve to ensure consistency with the departmental level. This should not involve too many risks, for the following reasons:

- a certain deterioration of performance in relation to the departmental estimates is acceptable, especially since the departmental estimates should be of good quality;
- the data from the income tax files should be quite useful;
- trend estimation and calibration on estimates at higher geographic levels (in this case the departmental estimates) both act as safeguards.

Of course, there is nothing prohibiting the use of the system to produce estimates for other sub-departmental geographic units.

At the departmental level, it does not seem useful to adapt the parameters (initial weights and norms) to population size; on the other hand, for sub-departmental levels, such an adaptation appears essential. Otherwise we run the risk of being much too strict for small areas. It would seem that a norm function of the following type might be appropriate:

$$NO_S = \alpha P^{\beta},$$

where NO_S is the norm for source S, P is the population of the area and α and β are two parameters that hypothetically depend on source S. The parameter β is obviously negative. If β equals -0.25, the norm doubles when the population is divided by 16. It also appears that the type of geographic area has an effect: the unexplained portion (le flou) would on average be greater for a commune of 50,000 inhabitants than for an employment area of the same size. The parameters α and β must be defined for each sub-departmental source, and where applicable, for each type of area.

8.2 Timetable

The greater the number of sources, the better the system functions. However, for a given year, data from the different sources become available at different times. Since the system is able to function with a variable number of sources, one can develop, at least at the departmental level, several sets of estimates for January 1 of year n: for example, interim estimates in the third quarter of year n, based on the first sources available, then semi-definitive estimates in the third quarter of year $n + 1$, based on more sources, and then final estimates in the third quarter of year $n + 2$. Different factors must be taken into account: the complexity of an operation, and the magnitude of the changes due to the addition of a source. It will be possible to assess the latter factor by simulations on the first years of implementation of the system.

8.3 Integration of an Additional Source

The system is flexible and modular. Therefore, integrating a new source into it does not pose any particular problem. It is merely a matter of determining the method to be used in order to obtain a good estimate of the net migration rate for each area. The range of methods envisaged by the team is large enough that in most cases, it should be possible to find a type of method that is appropriate to the source.

To determine the parameters (initial weight and norm) to be assigned to the new source in the synthesis, we suggest putting the system through a dry run, with parameters set arbitrarily but reasonably; it is obviously wise to start with a fairly high norm and a fairly low weight. By analysing the differences obtained between the net migration rates obtained from the new source and the synthetic rates, a better norm can be determined. The weight can then be adapted accordingly, using (for lack of anything better) an assumed relationship of quasi-proportionality between the weight and the inverse of the square of the norm. Obviously, this process can be iterated, with the parameters of the other sources also being changed as required. However, the tests conducted at the departmental level on the period 1982-1990 appear to show that the overall performance of the system is not highly sensitive to changes – even sizable ones – in the initial weights; it is therefore not necessary to determine these weights with great precision – nor, indeed, is it possible to do so – before the next census.

9. CONCLUSION

The "multi-source" population estimation system presented here is robust and flexible, without being overly complex. It can function with a variable number of sources. To integrate a new source into it, no long historical observation period is required. Aberrant data are detected automatically and corrected, so that they do not distort the estimates. The experiments carried out, while still not numerous, indicate that this system is effective. After a debugging and break-in period, it should be possible to use the system in production without too many risks pending the results of the next population census, planned for 1999.

ACKNOWLEDGEMENTS

This article results from the thinking and efforts of a team, led by the authors, which consisted of: Xavier Berne, Michel David, Michel De Bie, Sophie Destandau, Jacques Leclercq, Françoise Lemoine, Catherine Marquis and Marc Simon. The team benefited from the assistance of several departments of INSEE. The Statistical Methods Unit and its chief, Jean-Claude Deville, deserve special mention. The authors also wish to thank Philippe Ravalet for his contribution to the theoretical aspect of this article, as well as the editorial staff of *Survey Methodology* and the members of the editorial jury for their constructive comments.

REFERENCES

DECAUDIN, G., and LABAT, J.-C. (1996). Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population. Document de travail de méthodologie statistique, n° 9601. INSEE. Paris.

DESCOURS, L. (1992). Estimation de populations locales par la méthode de la taxe d'habitation. *Actes des Journées de méthodologie statistique,* 13 and 14 March 1991. INSEE. Paris.

GUÉGUEN, Y. (1972). Estimation de la population des villes bretonnes au 1.1.1971. *Sextant*, n° 4. INSEE. Rennes.

de GUIBERT-LANTOINE, C. (1987). Estimations de population par département en France entre deux recensements. *Population*, 6, 881-910.

HOAGLIN, D.C., MOSTELLER, F., and TUKEY, J.W. (1983*).* *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley.

LAURENT, L., and GUÉGUEN, Y. (1971). Essai d'estimation de la population des villes bretonnes. *Sextant*, n° 1. INSEE. Rennes.

LONG, J.F. (1993). Postcensal Population Estimates: States, Counties and Places. Population Division. Technical Paper No 3. U.S. Bureau of the Census. Washington DC.

STATISTICS CANADA (1987). *Population Estimation Methods, Canada.* Catalogue No. 91-528E. Ottawa.