Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 99-106
Statistique Canada

Une méthode synthétique, robuste et efficace, pour réaliser des estimations locales de population en France

GEORGES DECAUDIN et JEAN-CLAUDE LABAT[1]

RÉSUMÉ

La France ne disposant pas de registres de population, les recensements de la population y constituent la base du système d'informations socio-démographiques. Cependant, entre deux recensements, l'actualisation de certaines données est nécessaire, notamment à un niveau géographique fin, d'autant plus que les recensements ont, pour diverses raisons, tendance à s'espacer. Une mission, dont l'objectif était de proposer un système améliorant substantiellement le dispositif d'estimations locales de population en vigueur, a été créée en 1993 au sein de l'Institut National de la Statistique et des Études Économiques. Elle s'est consacrée à une double tâche: réaliser une synthèse efficace et robuste des informations apportées par différentes sources administratives et mobiliser un nombre suffisant de «bonnes» sources. Le système «multi-sources» qu'elle a conçu et qui est présenté ici est souple et fiable, sans être trop complexe.

MOTS CLÉS: Estimations de population; fichiers administratifs; estimation robuste.

1. INTRODUCTION

En France, comme dans tous les pays ne disposant pas de registres de population, les recensements de la population sont la base du système d'informations socio-démographiques. Cependant, ce sont des opérations très lourdes qui, à l'heure actuelle, ne peuvent être réalisées plus fréquemment que tous les sept ou huit ans. Dans l'intervalle, l'actualisation de certaines données est donc nécessaire, notamment à un niveau géographique fin, d'autant plus que les recensements ont, pour diverses raisons, tendance à s'espacer. Ainsi les estimations locales de population constituent un enjeu important pour l'Institut National de la Statistique et des Études Économiques (INSEE).

Malgré les progrès accomplis dans ce domaine, la situation, en 1993, pouvait paraître encore assez peu satisfaisante. Par rapport au recensement de la population de 1990, les estimations de population réalisées, sur la base du recensement précédent (1982), pour les départements métropolitains avaient présenté des écarts parfois importants.

L'INSEE a donc créé une mission à caractère méthodologique, chargée de proposer un système améliorant substantiellement le dispositif en vigueur. Initialement, le prochain recensement devait avoir lieu en 1997. Il semblait donc raisonnable de faire fonctionner le nouveau système de façon expérimentale jusqu'au recensement, afin de vérifier ses performances, avant de l'utiliser en production. Le report du recensement à 1999 a renforcé la nécessité d'aboutir vite, afin de pouvoir utiliser le nouveau système dès 1996.

Pour atteindre son objectif, la mission s'est consacrée, avec le maximum de pragmatisme, à une double tâche: réaliser une synthèse efficace et robuste des informations apportées par différentes sources administratives et mobiliser un nombre suffisant de «bonnes» sources. Le système «multi-sources» qu'elle a conçu, et qui est présenté ici, n'est pas trop complexe et semble efficace. On en trouvera une présentation plus détaillée dans Decaudin et Labat (1996).

2. PRINCIPALES CONCLUSIONS

Les principales conclusions de la mission sont les suivantes:

(1) Il est impossible d'améliorer les estimations de population totale au moyen d'enquêtes par sondage, à moins d'imaginer une enquête d'une taille telle qu'elle s'apparenterait à un recensement.

(2) Aucune source de données administratives ne reflète suffisamment bien les évolutions de population. Toutes les sources peuvent présenter localement des dérives, des ruptures, des à-coups..., qui ne sont pas toujours faciles à déceler. En outre, il est souvent très difficile, voire impossible, d'obtenir de l'organisme responsable, même à l'échelon local, des éléments d'explication et surtout, lorsqu'il s'agit d'une erreur, les éléments de correction. De toute façon, il est imprudent de se fonder sur une seule source administrative, aussi bonne soit-elle, car sa pérennité n'est jamais assurée.

(3) En revanche, il est possible d'améliorer substantiellement les estimations de population totale en utilisant simultanément plusieurs sources. Un système «multi-sources», analogue à celui présenté ici mais plus rudimentaire, a été testé rétrospectivement, sur la période intercensitaire 1982-1990, pour les 96 départements métropolitains. L'erreur moyenne (moyenne des écarts relatifs en valeur absolue avec les résultats du recensement de mars 1990) est descendue au-dessous de 0,9 %, alors que l'erreur moyenne commise à l'époque, avec le système d'estimation en vigueur, était de 1,4 %.

[1] Georges Decaudin et Jean-Claude Labat, Institut National de la Statistique et des Études Économiques, 18, Boulevard Adolphe-Pinard, 75675 Paris, CEDEX 14.

3. UTILISATION SIMULTANÉE DE PLUSIEURS SOURCES

Pour utiliser conjointement plusieurs sources, différentes méthodes sont envisageables.

Une méthode universelle – et simple à mettre en œuvre – est la *régression multiple*. Sous forme simplifiée, cela revient à utiliser, pour toute zone z, la relation suivante:

$$P(n+1,z)/P(n,z) = c + \sum_S (k_S N_S(n+1,z)/N_S(n,z)),$$

où $P(n,z)$ est la population de la zone z au 1er janvier de l'an n, les $N_S(n,z)$ sont les effectifs provenant de chaque source S à la même date et les k_S des coefficients, qu'on estime par régression multiple sur une période passée. c est ici un terme constant qui ne sert qu'à la régression, le *calage* sur la population nationale permettant de corriger la dérive éventuelle.

Cette méthode est utilisée dans certains pays, le Canada et les États-Unis notamment (voir par exemple Statistique Canada 1987 et Long 1993). Néanmoins, elle n'a pas été retenue car elle présente de nombreux inconvénients:

- il faut pouvoir estimer les coefficients; c'est-à-dire disposer des données de chaque source sur une période passée assez longue;
- les coefficients peuvent évoluer avec le temps, sans qu'on puisse maîtriser cette évolution;
- comme on l'a déjà dit, les sources administratives sont, pour des raisons diverses (changements de réglementation, à-coups de gestion, erreurs...), sujettes à ce qu'on peut appeler des «anomalies». Pour chaque source S, l'importance de ces «anomalies» se reflète en partie dans le coefficient k_S, plus ou moins selon que leur effet à moyen terme a été plus ou moins grand sur la période d'étalonnage; mais les anomalies interviennent néanmoins dans les estimations avec le même poids que les «bonnes» données de la même source. Les estimations sont alors fortement perturbées.

Une autre méthode est celle dite «*composite*». Chaque source sert à estimer la population d'une ou plusieurs classes d'âge: la classe d'âge X bien couverte par la source, mais aussi parfois une autre classe présentant à coup sûr une évolution très voisine de celle de la classe X (par exemple les «30-45 ans», si X représente les «moins de 18 ans»). Il faut alors disposer d'indicateurs appropriés pour les autres composantes de la population et gérer correctement la consolidation de ces estimations «par parties».

Ce genre de méthode, utilisé aux États-Unis (Long 1993), nous a paru problématique, notamment à cause de la difficulté à traiter convenablement les «anomalies».

Le *système «multi-sources» proposé* repose sur une synthèse robuste d'estimations provenant des différentes sources. Il combine un raisonnement démographique et des techniques purement statistiques. Il s'inspire des expériences menées à la Direction régionale de Bretagne de l'INSEE, au début des années 1970 (Laurent et Guéguen 1971, Guéguen 1972). La défaillance de l'une des sources n'empêche pas un tel système de fonctionner, même si ses performances sont un peu dégradées.

4. UNE BASE DÉMOGRAPHIQUE

Le raisonnement démographique qui est à la base du système est élémentaire: en supposant connue la population totale $P(n)$ d'une zone au 1er janvier de l'an n, la population $P(n+1)$ de la zone au 1er janvier de l'an $n+1$ s'en déduit par ajout des deux composantes de la variation au cours de l'année n: l'excédent naturel (naissances moins décès) d'une part, et le solde migratoire (immigrants moins émigrants) d'autre part.

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

En France, l'excédent naturel est fourni annuellement au niveau communal par les statistiques de l'état civil. Si ces dernières ne sont pas encore disponibles sous forme définitive, ce qui est souvent le cas au troisième trimestre de l'année $n+1$, il est facile de les estimer avec une faible marge d'incertitude.

La seule inconnue est donc le solde migratoire sur l'année n: $SM(n) = I(n) - E(n)$ ou, ce qui est équivalent, le taux de solde migratoire $T(n) = SM(n)/P(n)$. En d'autres termes, estimer la population revient à estimer le solde migratoire depuis la dernière date où cette population est connue (ou supposée telle), et réciproquement.

En France, les soldes migratoires ont une importance non négligeable mais néanmoins modeste par rapport à d'autres pays, comme le Canada ou les États-Unis par exemple. En outre, ils présentent en général une certaine inertie, du moins à des niveaux géographiques relativement agrégés. Une façon d'apprécier l'influence de leurs variations, d'une période intercensitaire à la suivante, consiste à mesurer les erreurs qu'on aurait commises sur chaque période, si on avait estimé les populations en reconduisant les taux de solde migratoire annuels moyens de la période précédente. Sur la période 1982-1990, pour les départements (sans la Corse), l'erreur moyenne en fin de période (en 1990, au bout de huit ans) n'aurait été que de 1,3 %. Il n'était pas sûr, au démarrage de la mission, qu'on puisse atteindre une précision nettement meilleure. Toutefois, en 1975 comme en 1982, l'erreur moyenne qu'on aurait commise, avec la méthode tendancielle, aurait été beaucoup plus forte: 2,8 % et 2,7 % respectivement (sur sept ans). On peut donc penser que la période 1982-1990 a été exceptionnelle et qu'à l'avenir les inflexions redeviendront plus marquées.

5. DES ESTIMATIONS ISSUES DES DIFFÉRENTES SOURCES

On tire de chaque source, par une méthode appropriée, une estimation du taux de solde migratoire annuel de l'ensemble de la population. Les méthodes qui peuvent être utilisées dépendent des données disponibles.

Pour chacune des sources expérimentées et jugées «bonnes», au moins au niveau départemental, une méthode est proposée. Les cinq sources retenues sont les suivantes: taxe d'habitation; abonnés électriques; enfants bénéficiaires d'allocations familiales; statistiques scolaires; fichier électoral.

Les données relatives à la composition des foyers fiscaux, figurant dans les fichiers de l'impôt sur le revenu, constituent une sixième source qui devrait fournir de très bons résultats. Cependant, jusqu'à présent, ces données n'ont été analysées que pour quelques départements et la méthode d'utilisation n'est pas encore complètement définie.

Il est proposé en outre d'intégrer au système une estimation tendancielle du taux de solde migratoire.

Deux catégories de méthodes sont utilisées. La première concerne les sources relatives aux ménages; la deuxième celles portant sur des individus.

5.1 Sources relatives aux ménages

Certaines sources fournissent une information sur l'évolution du nombre de ménages. C'est le cas des sources «*taxe d'habitation*» (TH) et «*abonnés électriques*». La taxe d'habitation est un des quatre principaux impôts directs locaux. Comme son nom l'indique, elle s'applique aux logements occupés, selon des modalités différentes pour les résidences principales et les résidences secondaires. C'est la situation au 1er janvier de l'année d'imposition qui est prise en compte. Depuis les années 1980, la source TH est à la base des estimations départementales de population réalisées par l'INSEE (Descours 1992); la source «abonnés électriques» lui a été substituée au début des années 1990, en raison des perturbations provoquées par une modification du système de gestion qui s'est généralisée progressivement à tous les départements.

La méthode retenue pour utiliser ces sources est classique dans son principe. Elle conduit directement à une estimation de la population totale et comporte trois étapes principales:

(1) estimation du nombre de ménages;
(2) estimation de la taille moyenne des ménages et passage à l'estimation de la population des ménages;
(3) ajout de la population «hors ménages».

Dans la première étape, on suppose que le nombre de ménages évolue comme les données fournies par la source (nombre de résidences principales TH ou nombre d'abonnés électriques). La seconde étape est la plus délicate. Elle repose à la fois sur l'utilisation des statistiques de personnes à charge contenues dans les fichiers TH et sur une estimation, de nature tendancielle, de la taille moyenne des ménages.

Dans le système «multi-sources» proposé, on passe au taux de solde migratoire, pour confrontation avec les autres sources, à l'aide des statistiques de l'état civil (*cf.* section 4).

5.2 Sources relatives à des individus

Les autres sources utilisées portent sur des individus. Seule une certaine tranche d'âge X de la population est en général couverte convenablement. La méthode comporte alors deux étapes principales:

(1) estimation, à partir de la source, du taux de solde migratoire de la population d'âge X;
(2) passage au taux de solde migratoire de l'ensemble de la population.

La deuxième étape repose sur la relation statistique suivante, observée dans le passé, entre la variation, d'une période à l'autre, du taux de solde migratoire global (T) et celle du taux de solde migratoire pour la population d'âge X (TX):

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

où δ_X est un coefficient voisin de 1, dépendant de la tranche d'âge X. Cette relation est voisine de celle utilisée par de Guibert-Lantoine (1987) pour estimer la population à partir des statistiques scolaires.

Pour les tranches d'âge correspondant aux différentes sources utilisées, les valeurs, estimées par régression linéaire, du coefficient δ_X(+/−2 écarts-types) sont présentées dans les tableaux 1 et 2.

Tableau 1
Estimation de δ_X sur les départements, hors Corse, soldes internes

Période 1	Période 2	Âge en fin de période		
		0-19 ans	10-14 ans	35 ans ou plus
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)

Tableau 2
Estimations de δ_X sur le couple de périodes 1975-1982 et 1982-1990, hors Corse, soldes totaux

	Âge en fin de période		
	0-18 ans	9-15 ans	35 ans ou plus
Départements	0,65 (+/-0,11)	0,57 (+/-0,10)	1,22 (+/-0,16)
Département – zone d'emploi	0,65 (+/-0,04)	0,59 (+/-0,04)	1,17 (+/-0,06)

Quant à la première étape, elle dépend de la source:

Fichier électoral

Les migrations électorales annuelles pour la tranche d'âge retenue (les «30 ans ou plus») sont fournies directement par le fichier électoral géré par l'INSEE. On passe du taux de solde migratoire électoral au taux de solde migratoire résidentiel en divisant le premier par un coefficient reflétant l'ampleur de la révision électorale.

Statistiques scolaires

Le solde migratoire des «5-9 ans» est obtenu en soustrayant leur effectif l'année n de celui des mêmes

générations l'année d'après (c'est-à-dire de l'effectif des «6-10 ans» l'année $n + 1$) et en défalquant les décès.

Enfants bénéficiaires d'allocations familiales

L'effectif des «0-17 ans» est estimé en supposant qu'il évolue comme le nombre d'enfants bénéficiaires d'allocations familiales. On en déduit un solde migratoire de «jeunes» en comparant cette estimation à l'effectif résultant d'une évolution sans migrations, c'est-à-dire sous le seul effet du mouvement naturel.

6. SYNTHÈSE

6.1 Principes

Les différentes estimations élémentaires du taux de solde migratoire annuel font l'objet d'un traitement statistique, afin d'en tirer un «taux synthétique», retenu comme estimation finale. Le traitement permet d'éliminer les valeurs aberrantes, de sous-pondérer les valeurs suspectes et, plus généralement, d'attribuer à chaque source un poids adapté à ses performances.

Plus précisément, chaque source pouvant «dériver», les différentes estimations élémentaires sont en général biaisées; on les corrige d'abord du biais national de la source correspondante pour l'année considérée, biais qu'on estime au préalable. En procédant ainsi, on suppose implicitement que l'écart entre le biais local et le biais national est de faible importance par rapport au flou irréductible. Lorsqu'on disposera d'estimations pour plusieurs années, on devrait pouvoir tester cette hypothèse, et, le cas échéant, la remplacer par une hypothèse mieux adaptée à la réalité, afin d'améliorer la correction des biais au niveau local.

Notons qu'une opération en apparence aussi simple que la correction du biais national nécessite néanmoins quelques précautions. La solution consistant à opérer un calage brutal sur le taux de solde migratoire national, considéré par définition comme la bonne référence, est peu satisfaisante, en raison des anomalies qui peuvent venir perturber le calage. Il est donc préférable d'estimer les biais au cours d'un processus où l'on élimine aussi les anomalies. Le processus est analogue à celui qui est utilisé pour la synthèse et qui est décrit ci-après. Cependant, la détermination des biais, supposés nationaux et donc calculés sur 96 départements, est moins sensible aux anomalies que celle des taux synthétiques, calculés sur un petit nombre de sources. Seules les anomalies importantes sont susceptibles de fausser sensiblement le calage des taux et doivent donc être corrigées.

Le taux de solde migratoire «synthétique» est une moyenne pondérée des estimations élémentaires ainsi «calées». On attribue à chaque source S un poids «a priori» W_S censé refléter sa précision à moyen terme. Mais de plus, pour une année et une zone données, ce poids est modulé pour prendre en compte le caractère plus ou moins vraisemblable du taux correspondant. Ainsi, un taux «anormalement éloigné» des taux issus des autres sources – en pratique d'une valeur centrale de l'ensemble des taux de la zone – voit son poids annulé ou réduit. Pour cela, on examine l'écart entre le taux provenant de chaque source et la valeur centrale retenue et on le compare à une «norme» d'écart NO_S propre à la source, déterminée empiriquement à partir des données disponibles: si l'écart est inférieur à «a fois» la norme, on ne modifie pas le poids a priori; s'il est supérieur à «b fois» la norme, on met le poids à 0; entre les deux, on multiplie le poids par un coefficient, compris entre 0 et 1, calculé par interpolation.

Notons que l'estimation tendancielle est formellement traitée comme celles provenant des sources exogènes; son poids est annulé lorsqu'elle est considérée comme non vraisemblable, parce que trop éloignée des autres estimations.

La synthèse est réalisée de manière automatique, ce qui assure une homogénéité et une logique explicite aux traitements mis en œuvre. Cela ne supprime pas, pour autant, la nécessité de contrôler les résultats obtenus.

6.2 Présentation théorique

Sur le plan théorique, on a cherché à utiliser les raisonnements et les techniques de l'estimation robuste, exposées par exemple dans Hoaglin, Mosteller et Tukey (1983). La méthode retenue s'inscrit dans le cadre des M-estimateurs de tendance centrale et plus précisément dans la catégorie des W-estimateurs, qui mettent en œuvre l'algorithme des moindres carrés repondérés.

Les taux de solde migratoire pour l'année n et la zone z issus des différentes sources S (et corrigés de leurs biais nationaux) étant notés $TC_S(n, z)$, le taux synthétique $T(n, z)$ est solution de l'équation implicite:

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

où la fonction Ψ est de type redescendant à point de rejet fini:

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

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

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

Un processus itératif permet d'affiner progressivement le traitement automatique des données suspectes.

6.3 Première analyse des distances de chaque taux à la valeur centrale des taux

(1) Pour chaque zone z, on calcule une première valeur centrale des taux «calés» $TC_S(n, z)$. La valeur centrale retenue doit être peu sensible à l'existence éventuelle de valeurs très éloignées pour certaines sources, mais aussi être d'autant plus influencée par une source que cette source est en moyenne plus précise. Dans ces conditions, plutôt que de choisir la médiane – qui répondrait à la première condition – on retient une

statistique de rang un peu plus élaborée, mais néanmoins simple, compte tenu du petit nombre de valeurs; cette statistique est la moyenne, pondérée respectivement par 1/2, 1/4, 1/4, des trois quartiles:

- la médiane des taux $TC_S(n,z)$ pondérés par les poids a priori W_S,
- le quartile inférieur (Q1) des taux pondérés,
- le quartile supérieur (Q3) des taux pondérés.

(2) Les taux $T1(n,z)$ ainsi obtenus sont calés sur le taux de solde migratoire du niveau supérieur, par simple translation:

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

où $P(n,z)$ est la population de la zone z au 1er janvier de l'an n et $TREF(n)$ le taux de solde migratoire du niveau supérieur (le taux national pour la synthèse départementale).

(3) On calcule, dans chaque zone, les écarts de chaque taux à cette valeur centrale calée:

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

(4) Pour chaque source et chaque zone, l'ampleur de cet écart est appréciée par rapport à la «norme» d'éloignement NO_S propre à la source. Cette «norme» est déterminée empiriquement à partir des données disponibles: c'est en principe la moyenne des écarts constatés dans le passé, anomalies exclues. Il en résulte une première modulation du poids affecté a priori à cette source:

- si $EC1_S(n,z) \leq a1\,NO_S$, où $a1$ est un paramètre à choisir (voisin de 2), on ne modifie pas W_S, poids a priori de S. Autrement dit, si $WM1_S(n,z)$ est le coefficient de modulation de W_S (coefficient compris entre 0 et 1), on prend $WM1_S(n,z) = 1$;
- si $EC1_S(n,z) > b1\,NO_S$, où $b1$ est un autre paramètre (voisin de 3), on met W_S à 0, c'est-à-dire qu'on élimine la source S: $WM1_S(n,z) = 0$;
- si $a1\,NO_S < EC1_S(n,z) \leq b1\,NO_S$, on interpole $WM1_S(n,z)$ en fonction de la valeur de $EC1_S(n,z)$:

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

(5) A l'issue de cette première phase, on dispose donc de nouveaux poids propres à chaque source et à chaque zone, qui permettent d'éliminer ou de sous-pondérer localement les taux suspects: $W1_S(n,z) = W_S WM1_S(n,z)$.

6.4 Itérations

(1) A l'aide des poids ainsi modifiés $W1_S(n,z)$, on estime pour chaque zone une nouvelle valeur centrale, en prenant cette fois la moyenne pondérée des taux:

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

(2) On cale chaque taux $T2(n,z)$ sur le taux de solde migratoire du niveau supérieur, par translation. On obtient $TC2(n,z)$.

(3) On calcule, dans chaque zone, les écarts de chaque taux au taux moyen calé: $EC2_S(n,z) = \mid TC_S(n,z) - TC2(n,z) \mid$. À partir de ces écarts, on calcule de nouveaux coefficients de modulation des poids a priori, en utilisant des paramètres $a2$ et $b2$, pouvant être différents de $a1$ et $b1$ (inférieurs en principe). On obtient ainsi de nouveaux poids $W2_S(n,z)$ prenant mieux en compte les anomalies, car celles-ci ont été appréciées par rapport à une meilleure tendance centrale. Avec ces poids, on estime un nouveau taux synthétique $T3(n,z)$, que l'on cale sur le niveau supérieur pour obtenir $TC3(n,z)$.

(4) On répète les opérations du point 3) avec les mêmes paramètres $a2$ et $b2$. Les tests menés au niveau départemental sur 1982-1990 montrent que la convergence est en général rapide; les taux sont très souvent stabilisés à partir de la quatrième itération.

7. MISE EN ŒUVRE AU NIVEAU DÉPARTEMENTAL

Le système d'estimation qui vient d'être présenté dans ses grandes lignes – et qui est destiné à être utilisé de façon opérationnelle pour les années 1990 et suivantes – a été mis en œuvre par la mission pour l'année 1990 au niveau départemental, avec les cinq sources suivantes: taxe d'habitation (TH), abonnés électriques (EDF), allocations familiales (AF), statistiques scolaires (EN), fichier électoral (FÉ), plus l'estimation tendancielle (TEND).

La figure 1 illustre les résultats obtenus pour quelques départements. Le tableau 3 présente les valeurs des poids et des normes retenues pour faire fonctionner le système. Ce tableau présente également certaines statistiques provenant de la synthèse des taux de solde migratoire et portant notamment sur les écarts entre les taux issus de chaque source et les taux synthétiques.

Tableau 3
Mise en œuvre pour l'année 1990 au niveau départemental
Paramètres et statistiques

	TH	EDF	AF	EN	FÉ	TEND
Poids	115	100	80	70	80	100
Norme	0,15	0,17	0,19	0,20	0,19	0,12
Nombre de taux	96	96	89	96	94	96
Moyenne des écarts	0,55	0,14	0,30	0,19	0,14	0,13
Nombre de taux «aberrants»	37	2	17	3	1	6
Moyenne des écarts sans les taux «aberrants»	0,15	0,13	0,16	0,16	0,13	0,11

Nota: - Coefficients (a; b) appliqués aux normes: (2,5; 3,5) à la première itération, puis (2; 3).
- Les valeurs des écarts et des normes correspondent à des taux exprimés en %.
- Les écarts sont calculés par rapport aux taux synthétiques après trois itérations.
- Les taux «aberrants» sont ceux dont le poids est annulé après trois itérations.

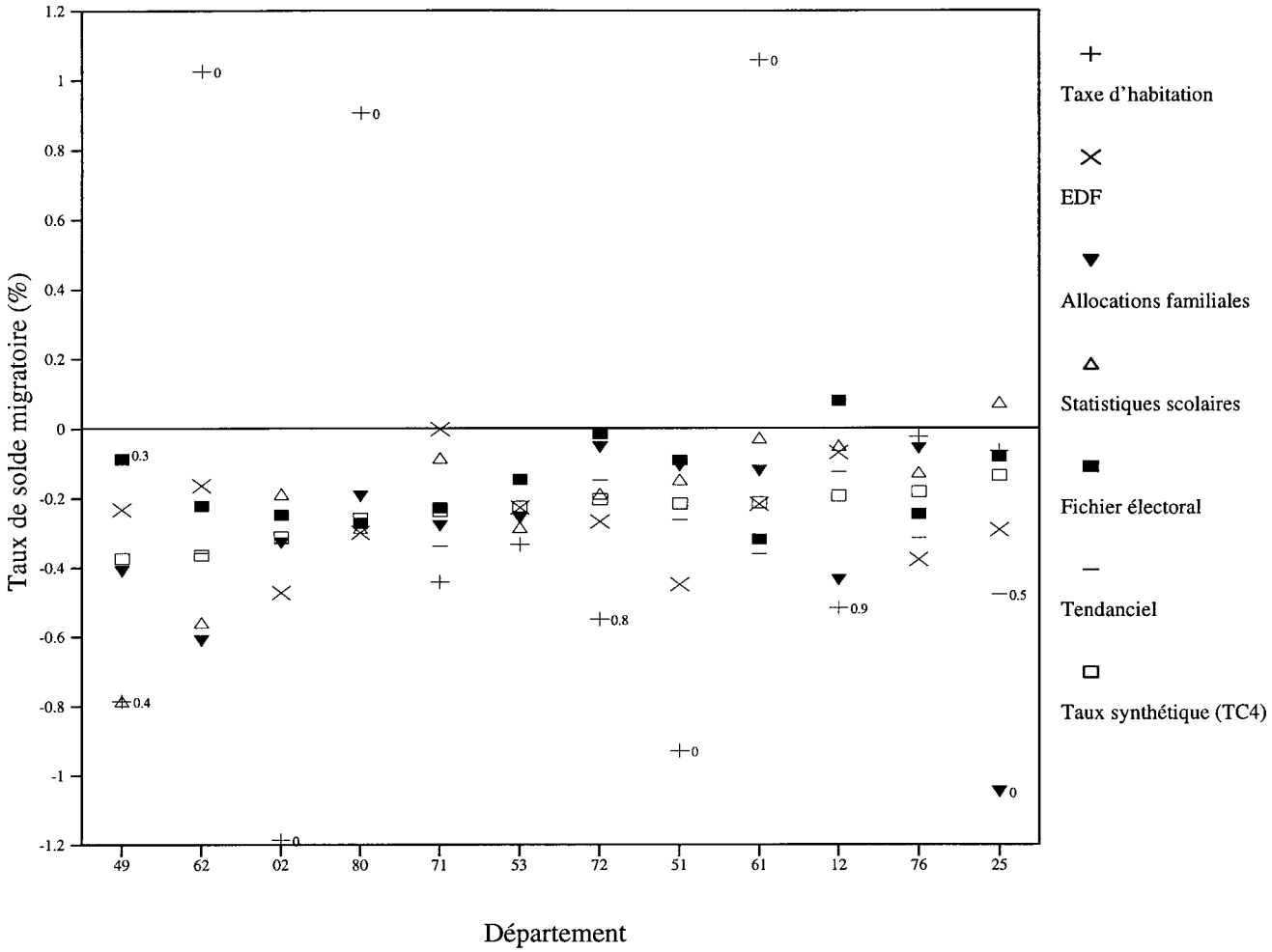

Figure 1. Synthèse des taux de solde migratoire de l'année 1990 pour douze départements, repérés par leur numéro (49, 62...). N.B.: TC4 est le taux synthétique obtenu après trois itérations. Lorsque le poids d'une source est annulé ou réduit, la valeur du coefficient de modulation (WM3) est indiquée.

Les résultats conduisent à penser que le système est encore plus efficace que ce qu'a indiqué le test rétrospectif sommaire réalisé sur la période intercensitaire 1982-1990 avec les mêmes sources. En effet, en dehors de la source TH, encore perturbée, les estimations provenant des différentes sources sont plus convergentes qu'elles ne l'étaient en moyenne dans le test rétrospectif (*cf.* tableau 4).

Cela n'a d'ailleurs rien d'étonnant, compte tenu du caractère rudimentaire du système testé sur la période intercensitaire 1982-1990. En effet les données utilisées étaient sommaires, voire fragmentaires, en raison de la difficulté à mobiliser en 1993 des données de gestion pour des années anciennes (1982, ...); en outre, les relations utilisées pour tirer de chaque source une estimation du taux de solde migratoire étaient simplistes; enfin, la méthode de synthèse était moins élaborée.

Notons que l'intégration d'autres sources, des données de l'impôt sur le revenu notamment, ne peut que renforcer encore l'efficacité du système.

Tableau 4
Moyenne des écarts dans le test rétrospectif

	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
Moyenne générale	0,43	0,30	0,41	0,39	0,32

Nota: - Le nombre de taux par année est généralement de 96, sauf pour AF (89) et FE (94).
- La source «fichier électoral» n'a pas fourni de taux pour 1986 ni 1987.
- La source «Taxe d'habitation» a commencé à être perturbée en 1987.
- Les valeurs des écarts correspondent à des taux exprimés en %.

8. COMPLÉMENTS

8.1 Niveaux infradépartementaux

L'utilisation de certaines sources peut devenir hasardeuse à un niveau géographique plus fin que le département, et cela pour différentes raisons: parce que les hypothèses sur lesquelles repose la méthode deviennent fragiles, parce que les effectifs sont faibles... Les statistiques scolaires sont notamment dans ce cas.

Cependant, on ne devrait pas courir trop de risques en faisant fonctionner le système pour les zones d'emploi; plus précisément pour les croisements «département * zone d'emploi» (environ 420 zones) permettant d'assurer la cohérence avec le niveau départemental.

En effet:

- on peut accepter une certaine dégradation des performances par rapport aux estimations départementales, d'autant que ces dernières devraient être de bonne qualité;
- les données tirées des fichiers de l'impôt sur le revenu devraient être d'un apport précieux;
- l'estimation tendancielle et le calage sur les estimations de niveau géographique supérieur (départementales en l'occurrence) jouent, l'une et l'autre, un rôle de garde-fou.

Notons que rien n'interdit, bien entendu, d'utiliser le système pour produire des estimations dans d'autres zonages infradépartementaux.

Au niveau départemental, il ne semble pas utile d'adapter les paramètres (poids «a priori» et normes) à la taille de la population; en revanche, pour les niveaux infradépartementaux, cette adaptation semble indispensable. Sinon on risque d'être beaucoup trop rigoureux pour les petites zones. Il semble qu'une fonction de norme du type suivant puisse convenir:

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

où NO_S est la norme de la source S, P la population de la zone et α et β deux paramètres dépendant a priori de la source S. Le paramètre β est évidemment négatif. Si β vaut $-0{,}25$, la norme double lorsque la population est divisée par 16. Il semble aussi que le type de zone intervienne: ainsi le flou serait en moyenne plus important pour une commune de 50,000 habitants que pour une zone d'emploi de même taille. Les paramètres α et β sont à définir pour chaque source infradépartementale et, le cas échéant, pour chaque type de zone.

8.2 Calendrier

Le système fonctionne d'autant mieux que le nombre de sources est plus important. Toutefois, les sources relatives à une même année sont disponibles de façon échelonnée dans le temps. Le système étant capable de fonctionner avec un nombre variable de sources, on peut élaborer, au moins au niveau départemental, plusieurs ensembles d'estimations au 1er janvier de l'an n: par exemple, des estimations provisoires au troisième trimestre de l'année n, à partir des premières sources disponibles, puis des estimations semi-définitives au troisième trimestre de l'année $n + 1$, assises sur davantage de sources et enfin des estimations définitives au troisième trimestre de l'année $n + 2$. Différents éléments sont à prendre en compte: la lourdeur d'une campagne, l'ampleur des modifications dues à l'ajout d'une source, ampleur qui pourra être appréciée par des simulations sur les premières années de mise en œuvre du système.

8.3 Intégration d'une source supplémentaire

Le système est souple et modulaire. L'intégration d'une nouvelle source ne pose donc pas de problème particulier. Il suffit de définir la méthode permettant d'en tirer une bonne estimation du taux de solde migratoire de chaque zone. La panoplie des méthodes envisagées par la mission est assez fournie pour que, dans la plupart des cas, on puisse y trouver un type de méthode adapté à la source.

Pour déterminer les paramètres (poids «a priori» et norme) à lui attribuer dans la synthèse, on suggère de faire fonctionner le système «à blanc» avec des paramètres fixés arbitrairement, mais de façon raisonnable; il est évidemment prudent de démarrer avec une norme plutôt forte et un poids plutôt faible. L'analyse des écarts obtenus entre les taux de solde migratoire issus de cette source et les taux synthétiques permet de déterminer une meilleure norme. On peut alors adapter le poids en conséquence, en se servant, faute de mieux, d'une relation supposée de quasi-proportionnalité entre le poids et l'inverse du carré de la norme. On peut évidemment itérer ce processus, en modifiant également, le cas échéant, les paramètres des autres sources. Toutefois, les tests réalisés au niveau départemental sur la période 1982-1990 semblent montrer que les performances globales du système sont assez peu sensibles à des variations, même assez importantes, des poids «a priori»; il n'est donc pas nécessaire de déterminer ces poids avec une grande précision, ce qu'on ne pourra pas faire, de toute façon, avant le prochain recensement.

9. CONCLUSION

Le système d'estimation de population «multi-sources» présenté ici est robuste et souple, sans être trop complexe. Il fonctionne avec un nombre variable de sources. On peut y intégrer une nouvelle source sans qu'il soit nécessaire de disposer d'une longue période d'observation rétrospective. Les données aberrantes sont décelées automatiquement et corrigées, de façon à ne pas perturber les estimations. Les expérimentations, encore peu nombreuses, qui ont été réalisées conduisent à penser que ce système est efficace. Après une phase de mise au point et de rodage, il devrait pouvoir être utilisé en production sans trop de risques, en attendant les résultats du prochain recensement de la population, prévu pour 1999.

REMERCIEMENTS

Cet article est le fruit des réflexions et des travaux d'une mission, animée par les auteurs, à laquelle ont collaboré: Xavier Berne, Michel David, Michel De Bie, Sophie Destandau, Jacques Leclercq, Françoise Lemoine, Catherine Marquis, Marc Simon. La mission a bénéficié de l'aide de différents services de l'INSEE. L'Unité «Méthodes statistiques» et notamment son chef, Jean-Claude Deville, méritent tout spécialement d'être cités. Les auteurs remercient également Philippe Ravalet pour son apport théorique, ainsi que la Rédaction de *Techniques d'enquête* et les deux arbitres pour leurs commentaires constructifs.

BIBLIOGRAPHIE

DECAUDIN, G., et 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 et 14 mars 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., et TUKEY, J.W. (1983*). Understanding Robust and Exploratory Data Analysis.* New York: John Wiley.

LAURENT, L., et 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.

STATISTIQUE CANADA (1987). *Méthodes d'estimation de la population, Canada.* N° 91-528F au catalogue. Ottawa.