UNE METHODE SYNTHETIQUE, ROBUSTE ET EFFICACE, POUR REALISER DES ESTIMATIONS LOCALES DE POPULATION

G. Decaudin et J.-C. Labat

1. Introduction

En France, comme dans tous les pays développés 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 96 départements métropolitains avaient présenté des écarts parfois importants. En outre, seules quelques Directions Régionales de l'INSEE faisaient des estimations infradépartementales, avec des succès incertains. Dans le Nord-Pas-de-Calais, la méthode d'estimation, fondée sur l'analyse du marché du travail masculin (Fontaine, 1986), avait donné de bons résultats mais ne pouvait pas être généralisée.

L'INSEE a donc créé une mission «Estimations localisées de population», chargée de proposer un système améliorant substantiellement le dispositif en vigueur. Depuis l'origine, cette mission s'est donné comme objectif non seulement de concevoir une méthode d'estimation, mais aussi de réaliser une «maquette» informatique permettant d'exploiter les données selon la méthodologie retenue. Initialement, le prochain recensement devait avoir lieu en 1997. Il semblait donc raisonnable de faire fonctionner ce 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.

2. Présentation générale du système d'estimation

2.1. 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 présentent en effet des «anomalies» locales : des dérives, des ruptures, des à-coups...Ces anomalies 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» relativement sommaire, a été mis en œuvre rétrospectivement, sur la période intercensitaire 1982-1990 (c'est-à-dire en fait pour les années 1982 à 1989), pour les 96 départements métropolitains. L'erreur moyenne, mesurée par la moyenne des écarts relatifs en valeur absolue avec les résultats du recensement de 1990 (erreur absolue moyenne en fin de période : EAM), est descendue au-dessous de 0,9 %. En comparaison, l'erreur moyenne commise à l'époque, avec le système d'estimation en vigueur, était de 1,4 %. En outre, la défaillance de l'une des sources n'empêche pas un tel système «multi-sources» de fonctionner, même si ses performances sont un peu dégradées.

2.2. Principes du système proposé.

Il n'est cependant pas simple d'utiliser conjointement plusieurs sources. Le système proposé repose sur la combinaison d'un raisonnement démographique et de techniques purement statistiques.

2.2.1. Une base démographique.

Le raisonnement démographique qui est à la base du système est élémentaire : en supposant connue la population totale d'une zone au 1er janvier de l'an n, la population 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. En France, comme dans tous les pays développés, l'excédent naturel est fourni par les statistiques de l'état civil, qui sont fiables et pérennes ; la seule composante à estimer pour obtenir la population au 1er janvier de l'an $n+1$ est donc le solde migratoire sur l'année 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.

2.2.2. Des estimations issues de différentes sources.

On tire donc 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 (section 3).

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 et abonnés électriques (section 4) ; enfants bénéficiaires d'allocations familiales (section 5.1) ; statistiques scolaires (section 5.2) ; fichier électoral (section 5.3).

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 exploitées que pour quelques départements. La méthode proposée devra donc être validée ou, le cas échéant, aménagée (section 5.4).

Il est proposé, en outre, d'intégrer au système une estimation tendancielle du taux de solde migratoire (section 6).

2.2.3. Synthèse.

Les différentes estimations du taux de solde migratoire annuel ainsi obtenues 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. 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.

Cette synthèse est réalisée de manière automatique, ce qui assure une homogénéité et une logique explicite aux traitements mis en œuvre (section 7). Cela ne supprime pas, pour autant, la nécessité de contrôler les résultats obtenus.

2.3. Mise en œuvre.

Le système a été utilisé, avec les cinq sources mentionnées (ainsi que l'estimation tendancielle), pour estimer les taux de solde migratoire départementaux de l'année 1990. Les résultats obtenus conduisent à penser qu'il est encore plus efficace que ce qu'a indiqué le test rétrospectif, réalisé avec les mêmes sources. Cela n'a d'ailleurs rien d'étonnant puisque la plupart des méthodes ont été sensiblement améliorées par rapport au test. Notons que l'intégration d'autres sources, des données de l'impôt sur le revenu notamment, ne peut que renforcer encore cette efficacité.

2.4. Niveaux géographiques infradépartementaux.

Afin de répondre au besoin d'estimations infradépartementales, on propose d'abord de mettre en œuvre le système pour les croisements «département * zone d'emploi», soit environ 420 zones. Bien entendu, seules les zones d'emploi (au nombre de 350 environ) présentent un intérêt. Le croisement proposé n'est qu'un zonage intermédiaire permettant d'assurer la cohérence avec le niveau départemental.

Toutefois, la zone d'emploi n'est pas, à la différence du département, un zonage universel. Pour réaliser des estimations par zone d'emploi il faut donc disposer de données administratives par commune, les 36 000 communes constituant en France les unités administratives de base. Or les sources utilisées ne sont pas toutes disponibles au niveau communal à partir du 1er janvier 1990. Les statistiques d'enfants bénéficiaires d'allocations familiales ne le sont que depuis le 31 décembre 1993.

En outre, 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 proposée 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 le zonage proposé. 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.

Bien que les arguments précédents soient encore largement valables dans ce cas, la proposition de réaliser également des estimations communales, calées sur le niveau «département * zone d'emploi», constitue à l'évidence un pari plus risqué. Toutefois, cette proposition a essentiellement pour objet de répondre, de façon simple, au besoin de disposer d'estimations pour des zones «sur mesure». Dès lors que ces zones ont une certaine taille, la précision des résultats devrait être acceptable. Notons que rien n'interdit, bien entendu, d'utiliser le système pour produire directement des estimations dans d'autres zonages emboîtés que ceux proposés.

2.5. Estimations par sexe et âge.

La répartition par sexe et âge de la population présente un grand intérêt pour de nombreux utilisateurs. On expose donc, en complément, différentes méthodes pour estimer cette répartition et on en propose une, simple à mettre en œuvre, pour le court terme. Cette méthode fournit une répartition cohérente de la population par sexe et âge pour les départements et les zones d'emploi (section 8).

2.6. 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. Les données définitives sur la composition des foyers

fiscaux au 1.1.*n* ne sont pas utilisables, en régime permanent, avant le deuxième trimestre *n*+2. Celles relatives à la Taxe d'habitation sont disponibles au deuxième trimestre *n*+1. Les quatre autres sources mentionnées le sont avec des délais beaucoup plus courts, de l'ordre de 6 à 8 mois.

Cependant, le système est capable de fonctionner avec un nombre variable de sources. Dans cette situation, on peut donc choisir d'élaborer, au moins au niveau départemental, plusieurs ensembles d'estimations au 1.1.*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 *n*+1, assises sur davantage de sources et enfin des estimations définitives au troisième trimestre *n*+2. Dans ce calendrier, on mènerait chaque année *n*, au troisième trimestre, trois campagnes d'estimations : d'abord la campagne définitive au 1.1.*n*-2, puis la campagne semi-définitive au 1.1.*n*-1 et enfin la campagne provisoire au 1.1.*n*. Différents éléments seront 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.

2.7. 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 est assez fournie pour que, dans la plupart des cas, on puisse y trouver un type de méthode adapté à la source. Ainsi, prenons le cas des «déclarations annuelles de données sociales» (DADS), que l'INSEE reçoit pour une très grande partie des salariés et qui sont désormais exploitées exhaustivement (à partir de celles relatives à 1992). Elles constituent a priori une source intéressante. La méthode adéquate pour les utiliser est vraisemblablement très proche de celle proposée pour le fichier électoral. Pour estimer le taux de solde migratoire en 1993 d'une zone donnée, on pourra procéder de la façon suivante : dans le fichier résultant de l'appariement des données de 1992 et de 1993 relatives à chaque salarié, on sélectionnera ceux dont on connaît la commune de résidence au 31.12.1992 et au 31.12.1993 ; au sein de cette population, on comparera les effectifs résidant dans la zone aux deux dates ; on en déduira un taux de solde migratoire de salariés pour 1993, qu'on devrait pouvoir assimiler à celui d'une certaine tranche d'âge (25-59 ans ?) ...

3. Méthodologie générale

3.1. Lien entre population et solde migratoire.

En supposant connue la population totale $P(n)$ d'une zone au 1er janvier de l'an n, la population $P(n+1)$ 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 $n+1$, 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 France, les soldes migratoires ont une importance non négligeable mais néanmoins modeste par rapport à d'autres pays, comme le Canada ou les Etats-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 intercensitaire 1982-1990, pour les départements (sans la Corse), l'erreur moyenne en fin de période (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.

3.2. Utilisation de sources administratives.

Il existe aujourd'hui un certain nombre de fichiers administratifs, potentiellement mobilisables, contenant chacun des informations sur la population - en fait sur des sous-populations particulières - ou sur des unités statistiques liées à la population (logements par exemple). Ces fichiers constituent le matériau de base de tout

système d'estimations localisées. Toutefois, la pérennité, ou la comparabilité dans le temps, de ces sources administratives n'étant jamais assurées, il est utile de développer des méthodes basées sur différents indicateurs, afin de pallier la défaillance éventuelle de l'un d'eux. De toute façon, l'utilisation simultanée de plusieurs sources, si elle est menée judicieusement, ne peut qu'améliorer la précision des estimations.

3.2.1. Données mobilisables.

Une source administrative est en général capable de fournir, chaque année n, pour chaque zone z d'un certain zonage (souvent pour chaque commune, donc aussi pour tout zonage supracommunal), l'effectif N(n,z) de la «population» concernée à une certaine date ; disons, pour simplifier, au 1er janvier. Le terme «population» est pris au sens large ; il peut s'agir d'individus, mais aussi de logements...

Beaucoup plus problématique est la fourniture directe des données suivantes (qui, naturellement, n'ont de sens que pour des individus) :

- les flux migratoires (internes) de chaque zone z à chaque zone i au cours de l'année n : $M(n,z,i)$;

- les soldes migratoires internes $SMI(n,z)$, qui sont tels que :

$$SMI(n,z) = \sum_i M(n,i,z) - \sum_i M(n,z,i) \ ;$$

- ou les soldes migratoires totaux $SM(n,z)$, tenant compte des immigrants en provenance de l'extérieur $IE(n,z)$ et des émigrants vers l'extérieur $EE(n,z)$:

$$SM(n,z) = SMI(n,z) + IE(n,z) - EE(n,z).$$

Cependant, en faisant intervenir les «créations» C(n,z) et les «disparitions» D(n,z) (entrées dans le champ et sorties du champ non dues aux migrations), la formule suivante relie les effectifs en n et n+1 :

$$N(n+1,z) = N(n,z) + SM(n,z) + C(n,z) - D(n,z). \qquad (1)$$

On peut ainsi obtenir indirectement le solde migratoire de la population couverte par la source ; à condition, bien entendu, de disposer des éléments C(n,z) et D(n,z), ou de pouvoir les estimer.

3.2.2. Utilisation symptomatique d'une source.

La méthode la plus simple, toujours applicable, pour utiliser une source est de considérer l'évolution de l'effectif N(n,z) comme un indicateur symptomatique de l'évolution de la population totale. De façon simplifiée, cela revient à supposer vérifiée la relation : P(n+1,z) = P(n,z) N(n+1,z) / N(n,z). En cas de dérive, phénomène courant, le calage sur la population nationale permet de corriger la composante nationale de cette dérive. Quant à la composante locale, on peut parfois supposer qu'elle évolue peu et reconduire celle observée dans le passé. Cependant, cette méthode simpliste donne en général de médiocres résultats.

3.2.3. Utilisation spécifique d'une source portant sur des individus.

Lorsque la source porte sur des individus, seule une certaine tranche d'âge *X* de la population est en général couverte convenablement. On peut alors estimer, à partir de la source, le taux de solde migratoire de la population d'âge *X*, puis en déduire une estimation du taux de solde migratoire de l'ensemble de la population.

1) Choix de la tranche d'âge.

Souvent le choix précis de la tranche d'âge *X* n'est pas évident. Dans ce cas, une analyse de corrélation, entre les évolutions fournies par la source d'une part et les recensements d'autre part, sur la dernière période intercensitaire, peut être utile. On retient alors la tranche d'âge qui donne la meilleure corrélation. Cela suppose évidemment que les données nécessaires soient disponibles.

Deux situations sont à distinguer :

(a) Les données de la source ne sont pas disponibles par génération (c'est-à-dire par année de naissance) : on ne peut faire varier la tranche d'âge que dans les données des recensements ;

(b) Les données de la source sont disponibles par génération : dans ce cas, évidemment plus favorable, on peut faire varier en même temps la tranche d'âge pour la source et les recensements ; la tranche d'âge X une fois choisie, on ne retient alors dans la source que les générations correspondantes. Notons que rien n'empêche de considérer plusieurs tranches *X1*, *X2*... ; cela présente l'avantage de fournir des estimations par tranche d'âge, mais l'inconvénient de compliquer la synthèse «multi-sources».

2) Estimation du taux de solde migratoire de la population d'âge X.

Si la source fournit directement des soldes migratoires, le problème est résolu. Il en est de même si on peut calculer des soldes migratoires au moyen de la relation (1).

Sinon on peut commencer par faire, à l'aide de la source, une estimation «symptomatique» de la population d'âge X. Cette estimation sera évidemment plus précise que celle de la population totale :

$$PX(n+1,z) = PX(n,z)\, N(n+1,z) / N(n,z)$$ dans la situation (a),

ou $$PX(n+1,z) = PX(n,z)\, NX(n+1,z) / NX(n,z)$$ dans la situation (b).

Là encore, le calage sur la population nationale d'âge X permet de corriger la dérive éventuelle. En comparant la population ainsi estimée à la population correspondante «attendue» en l'absence de migrations, on obtient, par différence, une estimation du solde migratoire. Notons que le calcul de la population «attendue» pose un problème, en général facile à résoudre, si on ne dispose pas d'estimation par âge annuel au 1.1.n.

3) Passage du taux de solde migratoire à l'âge X au taux de solde migratoire global.

J. Dekneudt (1990) a montré qu'en France il existait des relations, en général assez étroites :

- entre les taux de solde migratoire aux divers âges et le taux global ;

- mais aussi, ce qui est plus intéressant, entre les variations de ces taux d'une période intercensitaire à l'autre.

A l'époque l'analyse avait été faite au niveau régional. On pouvait craindre que les relations s'affaiblissent beaucoup au niveau départemental. On a donc repris la même analyse :

- par département, en utilisant les taux de solde migratoire internes fournis par les quatre derniers recensements (exploitation de la question sur la résidence antérieure) ;

- par département et par zone d'emploi, en utilisant les taux de solde migratoire totaux (obtenus par comparaison des effectifs aux recensements, génération par génération, en prenant en compte la mortalité) sur les deux dernières périodes intercensitaires.

De façon générale, on a confirmé la validité de la relation statistique suivante :

$$T(p2) = T(p1) + \delta_X \left(TX(p2) - TX(p1) \right),$$

où :

- $T(p1)$ et $T(p2)$ représentent les taux de solde migratoire annuels moyens sur deux périodes intercensitaires successives, $p1$ et $p2$;

- $TX(p1)$ et $TX(p2)$ représentent les mêmes taux pour la population d'âge X.

Pour les tranches d'âge correspondant aux différentes sources utilisées, les résultats obtenus sont très satisfaisants. Les valeurs estimées du coefficient δ_X (+/- 2 écarts-types) sont présentées dans les *tableaux 1 et 2*.

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

Période 1	Période 2	Age 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.

	Age 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)

3.2.4. Performances comparées de différentes méthodes.

Afin de comparer leur précision, différentes méthodes ont été appliquées aux statistiques scolaires sur la période 1982-1990 (94 départements, Corse exclue). L'erreur moyenne en fin de période (EAM) est :

- avec la méthode symptomatique, appliquée à l'ensemble des élèves du premier degré : de 2,8 % (avec calage national sur la population de 1990 supposée connue) ;

- avec la méthode appliquée par C. de Guibert-Lantoine (1987) à l'ensemble des élèves du premier degré (méthode correspondant à la situation (a) du § 3.2.3.) : de 2,4 % (résultat provisoire ; sur la période 1975-1982 l'EAM était de 2,2 %, sur 7 ans) ;

- avec la méthode correspondant à la situation (b) du § 3.2.3., appliquée par génération, c'est-à-dire la méthode proposée : de 1,9 % sans correction d'anomalies et de 1,6 % avec correction.

Ces résultats donnent une idée du gain de précision qu'on peut obtenir en affinant l'utilisation d'une source. Avec la dernière méthode, on aurait d'ailleurs abouti sans doute à un gain supérieur si la répartition par année de naissance n'avait pas été établie partiellement par sondage (cf. section 5.2).

3.2.5. Utilisation simultanée de plusieurs sources : régression multiple.

Une méthode universelle - et simple à mettre en œuvre - est la régression multiple. Sous forme simplifiée, cela revient à utiliser la relation suivante :

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

où les $N_s(n,z)$ sont les effectifs provenant de chaque source S 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 Etats-Unis notamment (voir par exemple Statistique Canada, 1987 et J.F. Long, 1993). Cependant, 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 ;

- les sources administratives sont, pour des raisons diverses (changements de réglementation, à-coups de gestion, erreurs...), assez souvent 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.

3.2.6. Utilisation simultanée de plusieurs sources : méthode «composite».

A partir de chaque source, on se contente d'estimer le solde migratoire (ou la population, ce qui est équivalent) de certaines classes d'âge : la classe d'âge X (cf. supra), mais aussi parfois une autre classe, non couverte par la source, mais 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 Etats-Unis (Long, 1993), nous a paru problématique, notamment à cause de la difficulté à traiter convenablement les «anomalies».

3.2.7. Utilisation simultanée de plusieurs sources : méthode proposée.

La méthode proposée, décrite en détail dans les sections suivantes, est largement empirique. Elle s'inspire des expériences menées à la Direction Régionale de Bretagne de l'INSEE, au début des années 1970 (L. Laurent et Y. Guéguen, 1971 ; Y. Guéguen, 1972). On estime, à partir de chaque source, un taux de solde migratoire global. On peut ainsi rapprocher ce dernier des estimations analogues tirées des autres sources, apprécier sa vraisemblance par rapport aux autres et faire assez facilement une synthèse de l'ensemble.

Il serait sans doute possible d'utiliser la méthode composite dans le même esprit. Cela suppose de pouvoir faire une synthèse robuste des indicateurs d'évolution fournis par les différentes sources, qui ne sont pas, pour la plupart, directement comparables les uns avec les autres. Ils reflètent en effet l'évolution : du nombre de résidences principales (taxe d'habitation) ; du nombre de résidences principales et secondaires (abonnés électriques) ; de la population de différentes tranches d'âge, pouvant se recouper ou non. Néanmoins, leur examen simultané devrait permettre d'apprécier leur cohérence et d'éliminer, ou de sous-pondérer, ceux qui paraissent peu vraisemblables compte tenu des autres. Avec ce genre d'analyse multidimensionnelle, on devrait aussi pouvoir traiter convenablement (sur un plan théorique) les situations de non-indépendance. Cette piste n'a pas été explorée. En effet, même si elle était viable et aboutissait à de meilleures estimations, elle conduirait sans doute à un système complexe et difficile à maîtriser. Or il est nécessaire que les utilisateurs régionaux et locaux puissent comprendre le système non seulement dans son principe, mais aussi dans son fonctionnement concret ;

c'est-à-dire puissent comprendre pourquoi et comment on aboutit à telle estimation finale avec telles données de base.

3.3. Apport d'enquêtes.

Ce point est mentionné essentiellement pour mémoire. En matière d'estimations localisées de population l'ordre de grandeur de précision d'une enquête est en effet très inférieur à ce qui est nécessaire et, en tout cas, à ce qu'il est possible d'obtenir par d'autres voies.

Pour estimer la population totale, une enquête de taille raisonnable ne peut pratiquement être d'aucune utilité, ni directement, ni même indirectement. Prenons l'exemple d'une enquête téléphonique destinée à estimer la taille moyenne des ménages (TMM) dans une zone de 100 000 habitants. Tout d'abord, même dans une enquête d'apparence aussi simple, les erreurs de mesure sur le nombre de personnes du ménage risquent d'être importantes ; mais surtout, pour estimer la TMM avec un écart-type de 1 %, il faut enquêter 3 300 ménages sur 40 000 (en supposant la variable «taille du ménage» de moyenne 2,5 et d'écart-type 1,5), ce qui représente un taux de sondage de 8 %. Un tel taux de sondage sur l'ensemble du territoire national est peu réaliste ; pourtant la précision théorique d'une telle enquête serait relativement médiocre : c'est, en moyenne, celle qu'on peut attendre, en France, d'une estimation tendancielle intelligente 4 ou 5 ans après un recensement.

De même, pour estimer la structure par sexe et âge, seule une enquête de taille prohibitive, s'apparentant à un recensement, pourrait apporter un gain de précision substantiel.

4. Taxe d'habitation, abonnés électriques

Disponibles au niveau communal, ces deux sources alimentent notamment la Banque de Données Locales de l'INSEE. La taxe d'habitation (TH) 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, cette source est à la base des estimations départementales de population réalisées par l'INSEE (L. Descours, 1992) ; la source «abonnés électriques» lui a été provisoirement substituée au début des années 1990 en raison des perturbations provoquées par une modification du système de gestion (procédure dite «IR-TH»).

Elles sont mises en œuvre de façon identique, en trois étapes :

- estimation du nombre de ménages ;

- estimation de la taille moyenne des ménages (TMM) et passage à l'estimation de la population des ménages ;

- ajout de la population «hors ménages».

Cette méthode générale peut s'appliquer à toute source reflétant l'évolution du nombre de ménages. Elle conduit directement à une estimation de la population totale. Dans le système «multi-sources» proposé, on passe au taux de solde migratoire, pour la confrontation avec les autres sources, à l'aide des statistiques de l'état civil.

4.1. Population «hors ménages».

Les erreurs commises sur la population «hors ménages» sont d'importance généralement faible, sauf éventuellement à un niveau local fin. C'est pourquoi l'hypothèse, admise jusqu'à présent, d'une stabilité par rapport au dernier recensement paraît acceptable, faute de mieux. L'idéal serait naturellement de constituer et de gérer annuellement un fichier des communautés, avec la population correspondante, ou au minimum la capacité d'accueil. Une solution intermédiaire consisterait à suivre seulement les établissements importants et à maintenir pour les autres l'hypothèse de stabilité.

4.2. Estimation du nombre de ménages.

Pour estimer le nombre de ménages, on suppose qu'il évolue comme le nombre de résidences principales TH ou le nombre d'abonnés électriques «domestiques et agricoles». En théorie, la notion de résidence principale TH et celle de résidence principale (ou de ménage) au recensement sont très proches, sinon identiques, ce qui ne signifie pas qu'en pratique il y ait coïncidence ... Cela confère en principe à cette source une supériorité sur la source «abonnés électriques», dont le principal inconvénient est de ne pas isoler les seules résidences principales dans l'ensemble des abonnements de particuliers. Lors de la synthèse, la source «abonnés électriques» pourra être affectée d'une sous-pondération spécifique, reflétant le poids des résidences secondaires dans la zone en 1990, et traduisant sa plus grande fragilité là où ce poids est élevé.

4.3. Estimation de la taille moyenne des ménages.

Il s'agit là du point le plus délicat de ce type de méthode : une erreur de 1 % sur l'estimation de la taille moyenne des ménages (TMM) se traduit par une erreur équivalente sur la population des ménages. Or, l'information manque sur l'évolution de la TMM à un niveau local.

Trois pistes ont été explorées :

- l'utilisation de résultats infranationaux issus des enquêtes sur l'emploi ;

- la modélisation des évolutions communales ;

- l'utilisation des statistiques sur le nombre de personnes à charge contenues dans les fichiers TH.

Après examen, c'est cette dernière piste que l'on propose de retenir.

4.3.1. Utilisation de résultats infranationaux issus des enquêtes sur l'emploi.

La méthode utilisée depuis quelques années pour estimer l'évolution de la TMM par département consiste à recourir aux résultats des enquêtes annuelles sur l'emploi par tranche d'unité urbaine (TUU en 10 tranches : 8 pour les communes urbaines, plus 2 pour les communes rurales) : l'indice annuel d'évolution par TUU est obtenu en exploitant le sous-échantillon commun à deux enquêtes successives. Le résultat est lissé par régression linéaire sur plusieurs années. Pour prendre en compte les disparités départementales d'évolution au sein de chaque TUU, un «différentiel» propre au croisement «département * TUU», calculé sur la dernière période intercensitaire, est ensuite appliqué à l'indice national de la TUU.

Or on constate, sur la période 1982-1990, que les évolutions de la TMM par TUU tirées des enquêtes sur l'emploi sont parfois sensiblement différentes de celles provenant de la comparaison des recensements, notamment dans les grandes agglomérations (*tableau 3*).

Tableau 3
Variation relative de la taille moyenne des ménages de 1982 à 1990 par TUU
Variation observée, variation estimée et écart-type théorique de l'estimation

En %

Taille d'unité urbaine (TUU) en 1982	Variation observée RP (recensements de population) (1)	Variation estimée EE (enquêtes sur l'emploi) (2)	Écart EE-RP (2)-(1)	Écart-type théorique de l'estimation EE
Rural profond	-5,6	-5,8	-0,2	1,0
Rural périurbain	-4,2	-3,2	+1,0	1,0
Unités de moins de 5 000 habitants	-6,1	-5,9	+0,2	1,4
Unités de 5 à 10 000 habitants	-6,5	-7,0	-0,5	1,5
Unités de 10 à 20 000 habitants	-6,8	-7,3	-0,5	1,6
Unités de 20 à 50 000 habitants	-7,2	-5,8	+1,4	1,5
Unités de 50 à 100 000 habitants	-7,1	-7,3	-0,2	1,5
Unités de 100 à 200 000 habitants	-7,5	-4,5	+3,0	1,7
Unités de 200 000 à 2 M. d'habitants	-6,2	-8,5	-2,3	1,0
Agglomération de Paris	-2,0	-3,9	-1,9	1,2
Ensemble	-5,3	-5,3	0,0	0,4

La précision théorique des évolutions de TMM par TUU estimées à partir des enquêtes sur l'emploi a été calculée par L. Meuric (1995) (cf. *tableau 3, dernière colonne*) : il en ressort que les écarts constatés sont compatibles avec les seuls aléas dus au sondage. Ces résultats inclinent à penser que la précision des estimations par TUU fondées sur les enquêtes sur l'emploi est insuffisante. On propose donc d'abandonner cette voie.

4.3.2. Estimation économétrique de l'évolution de la TMM.

On a tenté une modélisation de l'évolution communale de la TMM. Plus précisément on a cherché à expliquer les spécificités communales d'évolution de la TMM (définies en rapportant les évolutions communales à l'évolution nationale), à partir

de variables de structure, mesurées au dernier recensement, et de variables de flux connues (ou pouvant être estimées) annuellement. Ainsi, on a retenu les variables explicatives candidates suivantes (toujours rapportées aux moyennes nationales correspondantes) :

- poids des 15-19 ans et des 60-74 ans au dernier recensement ;

- taux annuel de natalité ;

- taux annuel de mortalité ;

- taux annuel d'évolution du nombre de résidences principales (la variable instrumentale pouvant être construite à partir de la source TH) ;

- niveau de la TMM au dernier recensement ;

- évolution de la TMM au cours de la dernière période intercensitaire.

Les ajustements ont été réalisés après stratification des communes. Cette stratification a été déterminée de façon empirique, en fonction de divers critères (taille des communes, appartenance à un ensemble urbain, ville-centre d'une agglomération, appartenance à l'Ile-de-France). On a ainsi défini 14 classes, dotées chacune d'un nombre suffisant de communes, et, pour chaque classe, on a ajusté un modèle annuel sur 1975-1982 d'une part, sur 1982-1990 d'autre part.

Utilisés pour estimer la TMM sur leur période d'étalonnage, ces ensembles de modèles conduisent à un gain de précision de 25 à 30 %, par rapport au maintien des spécificités communales d'évolution antérieures. Malheureusement, la situation se dégrade très fortement lorsqu'on les utilise sur la période suivante. Avec les équations étalonnées sur 1975-1982, la précision de l'estimation de la TMM en 1990 devient en effet un peu inférieure à celle de l'estimation obtenue par la méthode tendancielle.

Diverses tentatives d'amélioration ont été menées, mais n'ont pas conduit à des progrès décisifs. Cette modélisation, qui présente de surcroît l'inconvénient d'une certaine complexité de mise en œuvre, a donc été abandonnée.

4.3.3. Utilisation des données sur le nombre de personnes à charge contenues dans les fichiers TH.

En définitive, on propose d'utiliser l'information annuelle sur le nombre de personnes à charge contenue dans les fichiers TH. En effet, l'évolution du nombre moyen de ces personnes à charge par résidence principale est assez bien corrélée, sur la période 1982-1990, avec l'évolution du nombre de «0-19 ans» par ménage ;

au niveau départemental on obtient un coefficient de corrélation linéaire R = 0,75 ; de plus, la qualité de cette variable devrait s'améliorer avec la nouvelle procédure de gestion («IR-TH») mise en place dans les services fiscaux et consistant à rapprocher les données TH et les déclarations de revenu. On peut alors penser à décomposer le nombre moyen de personnes par ménage en trois composantes :

- le chef de ménage ;

- le nombre moyen d'enfants de moins de 18 ans ;

- le reste.

Par définition, la première composante est égale à 1. Faute d'information particulière, on fait évoluer tendanciellement la composante «reste», qui représente en moyenne le tiers de la TMM. Plus précisément, les tests menés sur 1982-1990 (avec l'évolution moyenne 1975-1982) montrent que l'on gagne en précision en «atténuant» l'indice tendanciel avant de l'appliquer : cette atténuation consiste à réduire l'écart entre l'indice tendanciel communal et l'indice tendanciel départemental, d'autant plus fortement que la commune est moins peuplée, donc la tendance passée plus fragile.

Quant à la composante «jeunes» de la TMM, on la fait évoluer en principe comme le nombre de personnes à charge par résidence principale TH. Toutefois, l'indice d'évolution correspondant peut avoir une valeur aberrante, par exemple dans la phase de mise en place de la procédure IR-TH, mais aussi à l'occasion d'autres perturbations administratives. On n'accepte donc cet indice que s'il est plausible. En pratique, la décision est prise en fonction :

- de l'éloignement par rapport à l'évolution tendancielle 1982-1990 d'une part ;

- et de la cohérence temporelle sur trois années successives d'autre part.

Le deuxième critère permet d'accepter des évolutions «éloignées» du tendanciel, à condition qu'elles soient compatibles avec les évolutions des deux années antérieures, d'où une présomption de non-anomalie de la source.

Comme pour la composante «reste», l'indice tendanciel de référence n'est pas directement l'indice moyen 1982-1990, mais un indice «atténué».

Cette méthode d'estimation des TMM communales a été testée sur la période 1982-1990, dans une version un peu moins élaborée : pas d'atténuation des évolutions tendancielles, contrôle grossier de la validité de l'information TH sur l'évolution du nombre de personnes à charge par ménage. Comme le montre le *tableau 4*, c'est elle

qui conduit aux estimations de la TMM les plus précises pour les niveaux géographiques considérés.

Tableau 4
Ecart-type des écarts relatifs signés (en %) sur la TMM par rapport au recensement de 1990 (Corse exclue)

Méthode d'estimation de la TMM	Départements	Zones d'emploi
Personnes à charge TH	1,11	1,38
Maintien des spécificités d'évolution 1975-1982 par «département * TUU»	1,24	....
Maintien des spécificités communales d'évolution 1975-1982	1,26	1,54
Econométrique (modèles étalonnés sur 1975-1982)	1,42	1,77
Enquêtes sur l'emploi par TUU + différentiels par «département * TUU»	1,46	....
Enquêtes sur l'emploi par TUU sans différentiels	1,86	....
Evolution communale uniforme (égale à l'évolution nationale 1982-1990)	1,88	2,07

Note : pour faire les agrégations, on a supposé connu le nombre de ménages en 1990.

5. Sources relatives à des individus

5.1. Enfants bénéficiaires d'allocations familiales.

Les statistiques établies par certains régimes d'allocations familiales fournissent le nombre d'enfants bénéficiaires au 31 décembre de chaque année. Ainsi, la Caisse Nationale des Allocations Familiales et la caisse centrale de la Mutualité Sociale Agricole publient régulièrement ces données, par caisse de gestion. Cela permet de disposer de données pour la quasi-totalité des départements. Malheureusement, aucune information ne peut être obtenue sur les changements de résidence, à quelque niveau que ce soit.

En ce qui concerne le régime «fonction publique» les informations analogues sont très difficiles à mobiliser. Quant aux autres régimes, ils sont globalement de moindre importance, mais peuvent avoir localement un poids non négligeable. La prise en compte des données analogues qu'ils pourraient fournir ne poserait aucun

problème. Il n'y a, en effet, pas de risque de doubles-comptes, chaque famille n'étant, en matière d'allocations familiales, affiliée qu'à un seul régime.

L'information sur le nombre d'enfants bénéficiaires d'allocations familiales des deux régimes considérés n'est a priori guère facile à exploiter pour les estimations localisées de population. Le champ couvert est en effet particulièrement complexe, en raison à la fois de l'existence d'autres régimes et des conditions d'attribution des allocations familiales (âge, nombre et situation professionnelle des enfants). Il peut varier si ces conditions changent ou si le domaine de compétence des deux régimes évolue.

Malgré ces restrictions, sur la période 1982-1990, les évolutions départementales du nombre de bénéficiaires sont bien corrélées avec celles du nombre d'enfants recensés ; c'est avec la population des «0-17 ans» que le coefficient de corrélation linéaire est le plus élevé : R = 0,93. Dans ces conditions, on propose d'exploiter cette source de la façon suivante :

1) Le nombre d'enfants de 0 à 17 ans est estimé par simple application de l'indice d'évolution du nombre d'enfants bénéficiaires. Selon toute probabilité, cette estimation s'écartera du total national issu du processus d'estimations nationales : sur la période 1982-1990, le nombre d'enfants bénéficiaires a crû en moyenne de 0,14 % par an, alors que celui des «0-17 ans» a diminué de 0,56 %. Il semble toutefois inutile de procéder à un calage : en effet, les taux de solde migratoire obtenus seront, comme ceux issus des autres sources, soumis à une procédure de détection et d'estimation d'un «biais» (cf. section 7.3).

2) Le solde migratoire de «jeunes» est obtenu en comparant l'effectif des «0-17 ans» au 1.1.n+1 ainsi estimé à celui résultant d'une évolution sans migrations (calculé en ajoutant les naissances de l'année *n* à l'effectif des «0-16 ans» au 1.1.*n* et en défalquant les décès en *n* des générations correspondantes).

Le taux de solde migratoire des «jeunes» TJ(n) est obtenu en rapportant le solde obtenu à la population correspondante, c'est-à-dire à l'effectif des «0-16 ans» au 1.1.n, auquel on ajoute les naissances de l'année n.

3) On passe de TJ(n) au taux de solde migratoire de la population totale, selon la méthode générale (cf. section 3.2.3) :

$$T(n) = TTL + 0{,}7\,(TJ(n) - TTLJ),$$

où *TTL* représente le taux de solde migratoire annuel moyen de la population totale entre 1982 et 1990 et *TTLJ* celui des «jeunes» sur la même période.

Au niveau communal, les statistiques d'enfants bénéficiaires d'allocations familiales des deux régimes considérés sont mobilisables à partir du 31 décembre 1993. Cependant, l'utilisation de ces données à ce niveau semble beaucoup plus hasardeuse qu'aux niveaux département et «département * zone d'emploi», notamment en raison du passage au taux de solde migratoire de la population totale. Il est possible également que l'analyse des données locales fasse apparaître des problèmes de fiabilité particulièrement aigus au niveau communal.

Le test réalisé sur la période intercensitaire 1982-1990, au niveau départemental, montre une assez bonne précision moyenne : l'écart-type des écarts relatifs signés sur la population totale est de 2,0 % sur 88 zones (sans la Corse et avec 7 départements regroupés pour l'Ile-de-France) ; l'erreur absolue moyenne est de 1,4 %. Quelques départements se trouvent cependant particulièrement mal estimés. C'est le cas notamment du Doubs, du Haut-Rhin et de la Haute-Savoie, où la méthode conduit à une nette sous-estimation (de -4 % à -8 %) : cela s'explique sans doute par un développement de l'emploi frontalier sur la période, en Suisse principalement, avec pour conséquence une proportion croissante d'enfants donnant droit à des prestations à l'étranger. S'il est impossible d'obtenir des informations permettant de quantifier ces phénomènes, on pourrait sous-pondérer la source dans les départements à forte proportion de travailleurs frontaliers.

5.2. Statistiques scolaires.

Jusqu'alors, les statistiques scolaires n'avaient jamais été utilisées à l'INSEE pour estimer la population. A l'INED, C. de Guibert-Lantoine (1987) les avait expérimentées sur la période 1975-1982 ; les résultats obtenus semblaient intéressants.

Au niveau départemental, on utilise les statistiques établies par le Ministère de l'Education Nationale au lieu de scolarisation. Ces statistiques portent sur la quasi-totalité des enfants scolarisés et sont disponibles par année de naissance. Toutefois, la répartition par année de naissance était, jusqu'à la rentrée scolaire de 1989, établie partiellement par sondage. Cela représentait, à l'évidence, un inconvénient pour réaliser des estimations localisées, compte tenu de l'importance relativement faible des effectifs concernés.

L'existence de données par âge d'une part et la possibilité d'en tirer facilement des soldes migratoires sont des facteurs favorables. Pour estimer le solde migratoire d'élèves, on compare dans une zone donnée, pour deux années successives, les effectifs d'un même ensemble de générations. Si l'on retient des générations pour lesquelles le taux de scolarisation est très proche de 100 %, la variation d'effectif représente le solde migratoire d'élèves, à l'effet près de la mortalité, connu par ailleurs.

On a exploité les données des rentrées scolaires n-1 et n pour les générations ayant de 4 à 13 ans révolus au 1.1.n. Les taux de solde migratoire 1982-1990 obtenus en chaînant les taux annuels ont été comparés au taux de solde migratoire intercensitaire des générations nées de 1974 à 1979. C'est en retenant, pour chaque couple d'années, les générations d'élèves ayant de 5 à 9 ans que l'on a obtenu la meilleure corrélation (coefficient de corrélation R = 0,92, sur 93 départements).

On propose d'appliquer la méthode suivante :

1) On assimile les effectifs inscrits à la rentrée n-1 par département de scolarisation à des effectifs au 1.1.n par département de résidence.

2) On estime, annuellement et par département, le taux de solde migratoire des «5-9 ans» : pour cela, on compare l'effectif des cinq générations d'élèves au 1.1.n+1 à l'effectif attendu en l'absence de migrations (effectif au 1.1.n moins décès en n) pour ces mêmes générations.

3) On passe du taux de solde migratoire des «5-9 ans» (TX) à celui de la population totale (T) à l'aide de la relation :

$$T(n) = TTL + 0{,}7\,(TX(n) - TTLX),$$

où TTL représente le taux de solde migratoire annuel moyen de la population totale entre 1982 et 1990 et TTLX celui des «5-9 ans» sur la même période.

Les statistiques annuelles par âge sont des statistiques au lieu de scolarisation. Bien qu'elles soient disponibles par commune et qu'aux âges envisagés les lieux de scolarisation et de résidence soient souvent proches, cela limite les possibilités d'utilisation à un niveau géographique très fin. Une solution alternative peut consister alors à estimer des effectifs d'élèves par âge et commune de résidence. Pour le premier degré, qui regroupe la quasi-totalité des élèves concernés, on dispose en effet d'une répartition des élèves par commune de résidence (mais sans répartition par âge). On peut toutefois craindre que la précision des estimations finales ainsi obtenues soit faible, même si la fiabilité des statistiques scolaires utilisées est bonne à un niveau fin.

5.3. Fichier électoral.

Les électeurs inscrits représentent une fraction très importante de la population : près des deux tiers de la population totale, environ 85 % de l'ensemble des «plus de 18 ans» (étrangers compris). Les taux d'inscription étant relativement faibles pour les premières années de la majorité, le rapport est encore plus élevé si on ne considère que les personnes de 30 ans ou plus.

Le fichier électoral est géré par l'INSEE, ce qui facilite son utilisation. Depuis 1988, les données de stocks sont mobilisables annuellement par commune d'inscription, sexe et âge. Les données de flux également, ce qui permet de calculer directement des soldes migratoires électoraux en fonction de ces variables. Il s'agit là d'un avantage majeur, car actuellement les fichiers permettant de rapprocher systématiquement les situations individuelles successives sont encore très rares.

Au niveau départemental, entre 1982 et 1990, les taux de solde migratoire électoraux reflètent très fidèlement les taux de solde migratoire résidentiels des Français. Pour la population de 35 ans ou plus en 1990, le coefficient de corrélation linéaire est de 0,98.

On propose d'appliquer la méthode suivante :

1) Le calcul des taux de solde migratoire peut être fait par sexe et par âge ; cependant, pour simplifier, on considère l'ensemble des personnes de 30 ans ou plus. On ne s'intéresse qu'aux personnes inscrites à la fois en début et en fin d'année. Soit NA30(n) le nombre de personnes considérées inscrites en n-1 (au 31 décembre) dans la zone et inscrites quelque part (n'importe où) en n. Soit NB30(n) le nombre de personnes considérées inscrites en n dans la zone et inscrites quelque part (n'importe où) en n-1.

Le solde migratoire électoral des «30 ans ou plus» au cours de l'année n est alors : $SM30(n) = NB30(n) - NA30(n)$; et le taux de solde migratoire : $T30(n) = SM30(n) / NA30(n)$.

2) Les taux de solde migratoire électoraux ainsi calculés dépendent, bien entendu, de l'ampleur de la révision électorale. Cette ampleur est très variable d'une année à l'autre. Elle dépend de l'importance attribuée par les électeurs potentiels aux élections de l'année suivante. L'utilisation annuelle de la source électorale suppose donc un redressement des taux de solde migratoire. On admet une relation de proportionnalité entre les taux observés et l'ampleur de la révision électorale. On obtient ainsi, pour chaque zone, un taux redressé en divisant les taux observés par un coefficient national $CORFE(n)$, indice de l'ampleur de la révision électorale : $TR30(n) = T30(n) / CORFE(n)$. Le calcul du coefficient $CORFE(n)$ ne nécessite pas une très grande précision. On retient comme base de détermination de l'ampleur le nombre de changements d'inscription des personnes de 30 ans ou plus. On propose de prendre comme ampleur moyenne pour 1991 la moyenne des sept années 1988 à 1994, et de la faire évoluer comme le stock d'électeurs de 30 ans ou plus en début d'année.

3) On passe du taux de solde migratoire $TR30(n)$ au taux de solde migratoire de l'ensemble de la population à l'aide de la formule suivante :

$$T(n) = TTL + 1{,}2\ (TR30(n) - TTL30),$$

où TTL représente le taux annuel moyen tous âges entre 1982 et 1990, et TTL30 le taux annuel moyen des «30 ans ou plus» sur la même période.

5.4. Impôt sur le revenu - données sur la composition des foyers fiscaux.

Les fichiers de l'impôt sur le revenu comportent des données sur la composition des foyers fiscaux. Ces données portent sur la quasi-totalité de la population et sont localisables par commune. Leur apport devrait être particulièrement important pour les estimations locales, là où les sources sont les moins nombreuses et les moins fiables.

Ils présentent cependant quelques inconvénients. Certaines tranches d'âge sont nettement moins bien couvertes que les autres, notamment la tranche «20-24 ans». Certaines personnes appartenant à un foyer fiscal peuvent résider ailleurs, par exemple les enfants à charge, notamment ceux qui poursuivent des études. Les fichiers sont disponibles tardivement : au printemps n+3 pour ceux, relatifs aux revenus de l'année n, fournissant la situation au 1.1.n+1. En outre, le rapprochement systématique des situations d'un même déclarant pour deux années consécutives, qui est réalisé notamment au Canada (Statistique Canada, 1995), est malheureusement impossible en France.

Les fichiers relatifs aux revenus des années 1989 à 1992 n'ont pas encore pu faire l'objet d'une exploitation systématique. La méthode suivante est donc proposée sous réserve de validation.

1) On calcule des populations fiscales communales, réparties par sexe et âge : NIR(n+1,j,x). Il faut notamment redresser certaines années de naissance non déclarées ou invalides (en faible proportion : environ 1 %) et procéder à une répartition par sexe des personnes à charge.

2) On corrige les populations fiscales communales par sexe et âge au 1.1.n+1 en les redressant par l'inverse des taux de couverture observés en 1990, supposés constants, soit :

$$NIRC(n+1,j,x) = NIR(n+1,j,x)\ COEFFIR(j,x),$$

avec :

$$COEFFIR(j,x) = P90(j,x) / NIR(90,j,x),$$

où P90 représente la population au 1.1.1990.

3) On suppose que la population fiscale corrigée NIRC(n+1,j,x) fournit une bonne estimation de la population réelle.

6. Prolongation tendancielle des taux de solde migratoire

Comme il a déjà été dit (section 3), les soldes migratoires, en France, évoluent en général avec une certaine inertie. La méthode consistant, faute de mieux, à estimer la population d'une zone en reconduisant chaque année le taux de solde migratoire annuel moyen de la dernière période intercensitaire n'est donc pas stupide. Les projections régionales de population sont d'ailleurs, le plus souvent, réalisées ainsi ; et, jusqu'à présent, les estimations départementales établies par l'INSEE intégraient, plus ou moins implicitement, une certaine dose de «tendanciel». L'idée de faire intervenir explicitement, dans la synthèse, une estimation purement tendancielle est donc venue naturellement. De cette façon on réduit très sensiblement le risque de produire une estimation «synthétique» aberrante ; tout particulièrement lorsque le nombre de sources disponibles est très faible. Un autre avantage est que le système peut fonctionner même si aucune source extérieure n'est encore disponible ; cela revient à faire une projection sur un an.

Dans le test rétrospectif réalisé sur la période 1982-1990, on a retenu simplement, comme estimation tendancielle, le taux de solde migratoire annuel moyen de la période
1975-1982. Pour la période 1990-1999, on propose d'appliquer une méthode plus élaborée reposant sur les considérations suivantes :

1) A un niveau infradépartemental, on a intérêt à ne pas appliquer brutalement le taux de solde migratoire annuel moyen de la période intercensitaire précédente. Les tests réalisés sur la période 1982-1990 montrent qu'on gagne en précision en «atténuant» le taux moyen de la période 1975-1982 avant de l'appliquer, et cela d'autant plus que ce taux est plus «atypique». Si le taux annuel moyen de la période précédente est T, il est préférable, en moyenne, d'utiliser un taux atténué TA :

$$TA = T_r + (T - T_r) / (1 + \lambda |T - T_r|) ,$$

où λ est un coefficient et T_r est un taux de référence tel que l'écart entre ce taux et le taux T (c'est-à-dire la valeur absolue de leur différence) permette d'apprécier le

caractère atypique de ce dernier. On obtient de bons résultats en prenant comme référence le taux départemental avec λ=50.

2) A tout niveau géographique, on a intérêt à prendre en compte l'évolution récente. Aux Pays-Bas, au niveau communal, la meilleure estimation par régression linéaire du taux de solde migratoire d'une année n à partir des taux des années précédentes ne fait quasiment intervenir que le taux de l'année n-1. Le taux de l'année n-2 intervient avec un coefficient relativement faible. Le taux moyen de la période correspondant à la dernière période intercensitaire française intervient de façon négligeable, sauf pour les deux ou trois premières années qui suivent.

En Belgique, la situation est assez différente de celle des Pays-Bas : le taux de l'année n-1 intervient avec un coefficient sensiblement plus faible et le passé plus ancien semble avoir un poids prédictif nettement plus important.

En France, il est évidemment impossible (tout au moins pour l'instant) de faire le même genre d'investigation. La méthode préconisée repose sur les considérations précédentes, en tenant compte du caractère imprécis des taux de solde migratoire estimés pour les années n-1, n-2...

La formule d'estimation du taux tendanciel avant calage est la suivante :

$$TTE(n) = t\,TTLA + t_1\,T(n-1) + t_2\,T(n-2) + t_3\,T(n-3), \qquad (2)$$

où T(n-k) est le taux de solde migratoire estimé pour l'année n-k. TTLA se déduit de TTL, taux de solde migratoire annuel moyen de la période 1982-1990, par la relation :

$$TTLA = TTLD + (TTL - TTLD)/\left(1 + 50\,|TTL - TTLD|\right),$$

où TTLD est le taux annuel moyen de la période 1982-1990 du département.

La méthode s'applique à chacun des trois niveaux géographiques de mise en œuvre de la méthode : département (dans ce cas les trois taux TTL, TTLD et TTLA sont égaux), «département * zone d'emploi», commune. Les coefficients proposés sont dans le *tableau 5*, pour les trois niveaux.

Tableau 5
Coefficients proposés pour la formule (2)

Année	*TTLA*	*T(n-1)*	*T(n-2)*	*T(n-3)*
1990	t=1	-	-	-
1991	t=0,7	t_1=0,3	-	-
1992	t=0,5	t_1=0,3	t_2=0,2	-
1993 & +	t=0,35	t_1=0,3	t_2=0,2	t_3=0,15

Chaque année n, on cale les taux TTE(n) sur le niveau géographique supérieur. Le calage est réalisé par simple translation, de façon à ajuster la moyenne pondérée des TTE(n) sur le taux estimé pour le niveau géographique supérieur TREF(n) (France, département ou «département * zone d'emploi», suivant le cas) : TCTE(n) = TTE(n) + (TREF(n) - TMOY(n)), où TMOY(n) est la moyenne des taux TTE(n) pondérés par les populations estimées au 1.1.n.

Pour chaque zone, c'est le taux calé TCTE(n) qui intervient dans la synthèse.

7. Synthèse des taux de solde migratoire

Le système d'estimation repose sur la synthèse des taux de solde migratoire issus de plusieurs sources, de manière à :

- accroître la fiabilité de l'estimation finale ;

- pouvoir continuer à fonctionner si une source devient défaillante ;

- permettre l'intégration de sources supplémentaires.

Cette synthèse est réalisée de manière automatique, ce qui assure une homogénéité et une logique explicite aux traitements mis en œuvre.

L'opération est réalisée pour chacun des trois niveaux géographiques proposés ; dans l'ordre : département, croisement «département * zone d'emploi», commune. Dans chaque cas, on utilise les taux élémentaires disponibles, qui varient suivant le niveau géographique et la date de réalisation. Les niveaux étant emboîtés, la cohérence d'un niveau par rapport au niveau supérieur est réalisée en fin d'opération par simple calage «descendant» des taux synthétiques : dans l'ordre, départements sur France,

croisements «département * zone d'emploi» sur départements, communes sur croisements «département * zone d'emploi».

7.1. Principes.

Chaque source pouvant «dériver», les estimations élémentaires provenant des différentes sources 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 (cf. section 7.3) ;

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

L'estimation tendancielle du taux de solde migratoire 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 *figure 1* illustre la synthèse des taux de solde migratoire départementaux de l'année 1990, réalisée à titre de démonstration (section 9, p. 400).

7.2. Détail de la méthode.

Sur le plan théorique, on a cherché à utiliser les raisonnements et les techniques de l'estimation robuste, exposées par exemple dans Hoaglin et al. (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 :

$$\sum_s W_s \,.\, NO_s \,.\, \Psi\left(\frac{TC_s(n,z) - T(n,z)}{NO_s}\right) = 0$$

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

$$\Psi(r) = r \qquad pour \quad |r| \le a$$

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

La synthèse étant la partie centrale du système, elle mérite d'être exposée en détail.

7.2.1. 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) On cale ensuite les taux $TI(n,z)$ sur le taux de solde migratoire du niveau supérieur, par simple translation :

$$TCI(n,z) = TI(n,z) + TREF(n) - \sum_z (TI(n,z)\,P(n,z)) \,/\, \sum_z P(n,z) \quad ,$$

où P(n,z) est la population de la zone z au 1.1.n, et TREF(n) le taux de solde migratoire du niveau supérieur (de la France métropolitaine pour la synthèse départementale).

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

$$EC1_S(n,z) = \left| \; TC_S(n,z) - TC1(n,z) \right| \; .$$

4) Pour chaque source et chaque zone, l'ampleur de cet écart est appréciée par rapport à une «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 $EC1_S(n,z) < a1\ NO_S$, 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 $a1\ NO_S \leq EC1_S(n,z) \leq b1\ NO_S$, 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 : $W1_S(n,z) = W_S\ WM1_S(n,z)$.

7.2.2. 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 :

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

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 : $EC2_S(n,z) = |TC_S(n,z) - TC2(n,z)|$ A 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.

7.2.3. Modulations spécifiques des poids pour certaines sources.

On a parfois de bonnes raisons de penser qu'une source donnée est a priori moins fiable dans certaines zones que dans d'autres. Dans ce cas, on propose d'introduire une modulation spécifique, de façon à la sous-pondérer localement. La différence avec les modulations décrites ci-dessus réside dans le fait que cette modulation spécifique est indépendante de la valeur du taux fourni par la source.

Ainsi, pour la source «abonnés électriques», on peut tenir compte du poids des résidences secondaires. Par exemple, à l'itération k :

$$Wk_{EL}(n,z) = W_{EL}\,WMk_{EL}(n,z)\big(1 - RSW90(z)\big),$$

où RSW90(z) est la part des résidences secondaires au recensement de 1990 dans la zone (calculée par rapport à l'ensemble des résidences principales et des résidences secondaires).

Pour la source «allocations familiales», on pourrait tenir compte du poids des résidents allant travailler à l'étranger, mal couverts par la source ; pour la source «fichier électoral», de la proportion d'étrangers...

7.3. Estimation du biais national de chaque source.

La synthèse décrite précédemment suppose que les taux élémentaires soient si possible sans biais. Les biais sont généralement faibles. Leur élimination présente

donc davantage d'importance à un niveau géographique agrégé (département notamment), où les taux de solde migratoire sont relativement peu élevés, qu'au niveau communal, où ces taux peuvent être d'ampleur bien plus grande.

C'est pourquoi il est proposé :

- de se limiter à l'estimation annuelle, pour chaque source, d'un biais national (en supposant que toutes les zones sont affectées du même biais) ;

- d'estimer ce biais à partir des seuls taux départementaux.

La solution simple consistant à opérer un calage brutal sur le taux national, considéré par définition comme la bonne référence, est peu satisfaisante. Dans ce cas en effet, toute anomalie d'un taux dans un département se répercute, via le calage, sur les autres départements. Il est donc préférable d'estimer les biais au cours d'un processus où l'on détecte aussi les anomalies. Cependant, la détermination des biais (supposés nationaux) ne nécessite pas une détection des anomalies aussi fine que la synthèse proprement dite. Seules les anomalies importantes sont susceptibles de fausser le calage des taux et doivent donc être corrigées.

7.3.1. Principe de la détection des taux en anomalie.

La détection des anomalies est menée chaque année, département par département. Comme on ne connaît pas la vraie valeur du taux de solde migratoire départemental, on prend comme estimation une valeur centrale robuste des taux issus des diverses sources ; robuste, c'est-à-dire beaucoup moins susceptible de fortes anomalies que chacun des taux pris séparément. Le caractère anormal d'un taux donné est alors apprécié en comparant sa distance à cette valeur centrale avec une distance considérée comme «normale», ou «habituelle». La valeur centrale retenue est la statistique de rang utilisée dans la première phase de la synthèse : la moyenne, pondérée respectivement par 1/2, 1/4, 1/4, des trois quartiles (médiane, Q1 et Q3) pondérés. Ces valeurs centrales, calculées à partir de taux non calés, sont elles-mêmes en général affectées d'un «biais» : leur moyenne pondérée diffère quelque peu du taux national. On les corrige toutes de cette différence : les inconvénients propres à ce calage sont ici atténués, en raison du risque moins grand de grosse anomalie sur ces valeurs centrales que sur les taux issus des diverses sources.

7.3.2. Principe de l'estimation du biais.

Si, pour une source donnée *S*, aucun des 96 taux départementaux ne se trouve en anomalie, alors, par hypothèse, le biais de la source s'estime simplement par la différence entre la moyenne pondérée des 96 taux et le taux national. Ou, ce qui est équivalent, par la moyenne pondérée des différences départementales entre taux et valeur centrale calée. Cette dernière formulation présente l'avantage de s'appliquer

lorsque certains des 96 taux manquent. En cas d'anomalies, le principe de la méthode consiste à l'appliquer en considérant les départements où S est en anomalie comme manquants.

On voit que l'efficacité de la méthode est basée en grande partie sur la détermination d'une «bonne» valeur centrale. Aussi est-il nécessaire de procéder de façon itérative, en affinant progressivement et de concert estimation de la valeur centrale, détection des anomalies et estimation du biais.

7.3.3. Un processus itératif.

1) Le début du processus est simple : le premier ensemble de valeurs centrales départementales calées est calculé comme indiqué en 7.3.1, en retenant toutes les sources disponibles. Pour chaque source S, la première estimation du biais est obtenue en retenant la médiane des 96 différences départementales entre taux issu de S et valeur centrale, plutôt que la moyenne pondérée. En effet, l'expérience menée sur les années 1982 à 1990 montre qu'en cas d'anomalies nombreuses, le processus de convergence est ainsi beaucoup plus efficace.

2) Disposant pour chaque source d'une première estimation de son biais, on corrige de ce biais chacun des taux départementaux correspondants. A partir de ces taux corrigés, on calcule pour chaque département une nouvelle valeur centrale calée. On procède alors à une première détection des sources en anomalie, en analysant, département par département et source par source, les écarts de chaque taux corrigé à cette nouvelle valeur centrale. Lorsque cet écart est, en valeur absolue, supérieur à «a NO_S», où NO_S est la «norme» utilisée dans la synthèse et a un paramètre à choisir (voisin du paramètre a1, c'est-à-dire, lui aussi, voisin de 3), on considère la source S comme étant en anomalie.

3) Ayant ainsi détecté, pour chaque source S, un premier ensemble de départements en anomalie, on peut estimer une nouvelle valeur du biais de S, qui remplace la valeur initiale. Pour cela, on affine au préalable la détermination des 96 valeurs centrales départementales, en excluant de leur calcul dans chaque département les sources en anomalie. Le nouveau biais de la source S est alors estimé par la moyenne pondérée des différences départementales entre taux et valeur centrale, en ne retenant dans cette moyenne que les seuls départements où S n'est pas en anomalie.

4) L'itération suivante consiste, source par source, à corriger de cette nouvelle estimation du biais les 96 taux départementaux correspondants. D'où calcul d'une nouvelle valeur centrale pour chaque département, à partir des taux corrigés non en anomalie ; et, pour chaque source, nouvelle phase de détection de départements en anomalie. Puis nouvelle estimation du biais de chaque source. Et ainsi de suite...

Les essais menés sur la période 1982-1990 ont montré qu'avec un coefficient a égal à 3 ou 3,5 le processus converge assez vite. Ils ont également montré que cette méthode automatique de calage peut fonctionner convenablement même en présence de nombreuses anomalies. Il est cependant indispensable d'en contrôler les résultats. Au cas où, pour une source, le biais estimé pour une année serait très différent des années précédentes, il est évident que des investigations ad-hoc seraient nécessaires avant de procéder à la synthèse des taux ;
en particulier, si le biais est élevé (nettement supérieur à 1 %), il faudrait être très vigilant, pour deux raisons :

- le processus peut ne pas bien fonctionner ;

- l'hypothèse d'un biais universel peut être très éloignée de la réalité.

8. Estimations par sexe et âge

La répartition par sexe et âge de la population présente un grand intérêt pour de nombreux utilisateurs. Il s'agit d'estimer l'effectif de sexe j et d'âge x de la zone z en n+1 (au 1er janvier de l'année n+1) : P(z,j,x,n+1). On suppose réalisée l'estimation de la population totale P(z,n+1) ou, ce qui est équivalent, celle du taux de solde migratoire global T(z,n) de l'année n.

Plusieurs méthodes sont envisageables.

8.1. Estimation par calage à l'aide du logiciel CALMAR.

Une méthode générale, pouvant s'appliquer à différentes structures, consiste à faire simplement une estimation par calage sur marges en utilisant le logiciel CALMAR (Sautory, 1993). On connaît en effet :

- la population de sexe *j* et d'âge *x* de chaque zone au dernier recensement ;

- la population nationale de sexe *j* et d'âge *x* en *n*+1 (1re marge) ;

- la population totale de chaque zone en *n*+1 (2ème marge).

La méthode consiste à chercher la répartition P(z,j,x,n+1) qui soit la plus «proche» de la structure initiale, tout en respectant les deux marges.

En procédant ainsi, on ne tient pas compte de l'effectif initial de la génération ayant l'âge x en n+1. Cependant, cet inconvénient théorique n'est peut-être pas très grave

en pratique. En effet les spécificités locales de structure par âge ont souvent tendance à «se perpétuer», en raison de l'inertie de celles des taux démographiques.

A titre expérimental, cette méthode a été appliquée pour estimer la répartition par grand groupe d'âges des départements en 1990, en partant de celle de 1982, les marges de 1990 étant supposées connues sans erreur. Les écarts avec le recensement de 1990 sont en moyenne les suivants (sur 94 départements - hors Corse) :

- moins de 15 ans : 3,0 %

- 15-34 ans : 1,5 %

- 35-64 ans : 1,6 %

- 65 ans ou plus : 2,9 %

Les résultats obtenus sont relativement bons pour les «15-34 ans» et les «35-64 ans». Il semble cependant préférable de procéder à une estimation directe des taux de solde migratoire par sexe et âge, puis de caler les estimations qui en résultent sur les mêmes marges : sur la population totale de chaque zone d'une part et sur la population nationale par sexe et âge d'autre part.

La phase cruciale est alors l'estimation des taux de solde migratoire par sexe et âge. La méthode suivante peut s'appliquer pour les départements et les croisements «département * zone d'emploi» :

8.2. Utilisation «à l'envers» de la relation statistique entre variation du taux de solde migratoire global et variation du taux à l'âge x.

On déduit du taux de solde migratoire global T(z,n) de l'année n une estimation TSY(z,j,x,n) du taux de solde migratoire du sexe j à l'âge x par la relation :

$$TSY(z, j, x, n) = TTL(z, j, x) + \Delta(j, x)\big(T(z, n) - TTL(z)\big),$$

où TTL(z) est le taux annuel moyen global de la dernière période intercensitaire et TTL(z,j,x) le taux analogue pour le sexe j et l'âge x. $\Delta(j,x)$ est un coefficient estimé à partir des variations de taux observées au cours des deux périodes 1975-1982 et 1982-1990.

La précision de cette estimation est liée à celle du taux de solde migratoire global (a priori assez bonne), mais aussi à la précision de la relation supposée. A titre

expérimental, la méthode a été appliquée sur la période 1982-1990, par département, en prenant comme références (TTL) les taux de la période 1975-1982 et en supposant les marges de 1990 connues sans erreur. Les écarts avec le recensement de 1990 sont en moyenne les suivants (sur 94 départements - hors Corse) :

- moins de 15 ans : 0,9 %

- 15-34 ans : 0,9 %

- 35-64 ans : 0,6 %

- 65 ans ou plus : 1,1 %

Les résultats sont nettement meilleurs qu'avec la première méthode. Il faut cependant garder à l'esprit qu'ils ne sont pas extrapolables sans précaution, puisque, dans les deux cas, les marges de 1990 ont été supposées connues.

8.3. Utilisation directe de certaines sources.

Certaines sources fournissent directement des informations sur certaines tranches d'âge :

- statistiques scolaires : 5-9 ans ;

- allocations familiales : 0-17 ans ;

- fichier électoral : 30 ans ou plus ;

- impôt sur le revenu : presque tous les âges.

En réalité, du fait des corrélations liant statistiquement les soldes migratoires aux divers âges, on peut dire que chaque source apporte directement ou indirectement des informations (plus ou moins fiables) sur chaque âge. Par exemple les statistiques scolaires devraient fournir une information d'assez bonne qualité sur les «35-44 ans».

Le cas d'une source fournissant un taux de solde migratoire pour une seule tranche d'âge est particulièrement simple : ainsi, pour la source «allocations familiales» (repérée par «AF»), les taux par âge détaillé peuvent être estimés par :

$$T_{AF}(z,j,x,n) = TTL(z,j,x) + C_{AF}(j,x)\,(TJ_{AF}(z,n) - TTLJ(z)).$$

Notons cependant qu'il faudrait sans doute «caler» les taux TJ_{AF} au préalable.

La façon de traiter une source S fournissant directement des taux de solde migratoire Tx_i pour différents âges x_i est moins évidente. Même pour estimer le taux relatif à un âge couvert par la source, on peut avoir intérêt à prendre également en compte les taux à d'autres âges fournis par la source. On pourrait imaginer une estimation par combinaison linéaire du type suivant :

$$T_S(z,j,x,n) = TTL(z,j,x) + \sum_i (Cx_i(j,x)(Tx_i(z,n) - TTLx_i(z)))$$

où les $Cx_i(j,x)$ seraient des coefficients, relatifs à la source S, dont la somme sur i serait voisine de l'unité. Ces coefficients seraient à déterminer. Les x_i pourraient être des groupes d'âges (pour restreindre le nombre de variables), mais il serait plus facile de calculer des taux par année d'âge.

Finalement, à chaque âge, le taux qu'on retiendrait pourrait être une moyenne pondérée des estimations fournies par chaque source, le poids de chaque source variant en fonction de l'âge. Ainsi le poids de la source scolaire serait a priori relativement fort vers 5-10 ans et vers 35-44 ans et très faible, voire nul, au-dessus de 60 ans. L'estimation TSY (cf. section 8.2) pourrait intervenir dans cette moyenne pondérée. Les poids seraient à définir, avec une part d'arbitraire assez grande.

Le risque peut provenir de la défaillance d'une source. Une synthèse par âge, analogue à celle réalisée pour l'estimation globale, ne paraît pas envisageable. On peut toutefois limiter les risques en prenant en compte les coefficients de modulation déterminés dans la synthèse globale ; ainsi une source dont le poids a été annulé dans cette synthèse n'interviendrait pas non plus dans les estimations par âge.

La piste, assez large, qui vient d'être ouverte n'a pas été explorée plus avant. Il est d'ailleurs possible que le seul apport de la source «impôt sur le revenu» permette d'améliorer substantiellement les estimations par âge. Dans ce cas, le pragmatisme pourrait conduire à se limiter, au moins dans un premier temps, à cette seule source.

9. Mise en œuvre

Le système d'estimation qui vient d'être présenté - 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 (FE), plus l'estimation tendancielle (TEND).

La *figure 1* illustre les résultats obtenus pour quelques départements. Le *tableau 6* présente les valeurs des poids et des normes retenues pour faire fonctionner le

système, ainsi que certaines statistiques issues de la synthèse des taux de solde migratoire, portant notamment sur les écarts entre les taux issus de chaque source et les taux synthétiques.

Tableau 6
Mise en œuvre pour l'année 1990 au niveau départemental

Paramètres et statistiques

	TH	EDF	AF	EN	FE	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	
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	

Notes : - Coefficients appliqués aux normes : *a1*=2,5 ; *b1*=3,5 ; *a2*=2 ; *b2*=3.
- Les valeurs des écarts et des normes correspondent à des taux exprimés en %.
- Les écarts sont calculés par rapport au taux synthétique après trois itérations (*TC4*).
- Les taux «aberrants» sont ceux dont le poids est annulé (*WM3*=0).

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 par la procédure «IR-TH», les estimations provenant des différentes sources sont plus convergentes qu'elles ne l'étaient en moyenne dans le test rétrospectif, comme le montre le *tableau 7*.

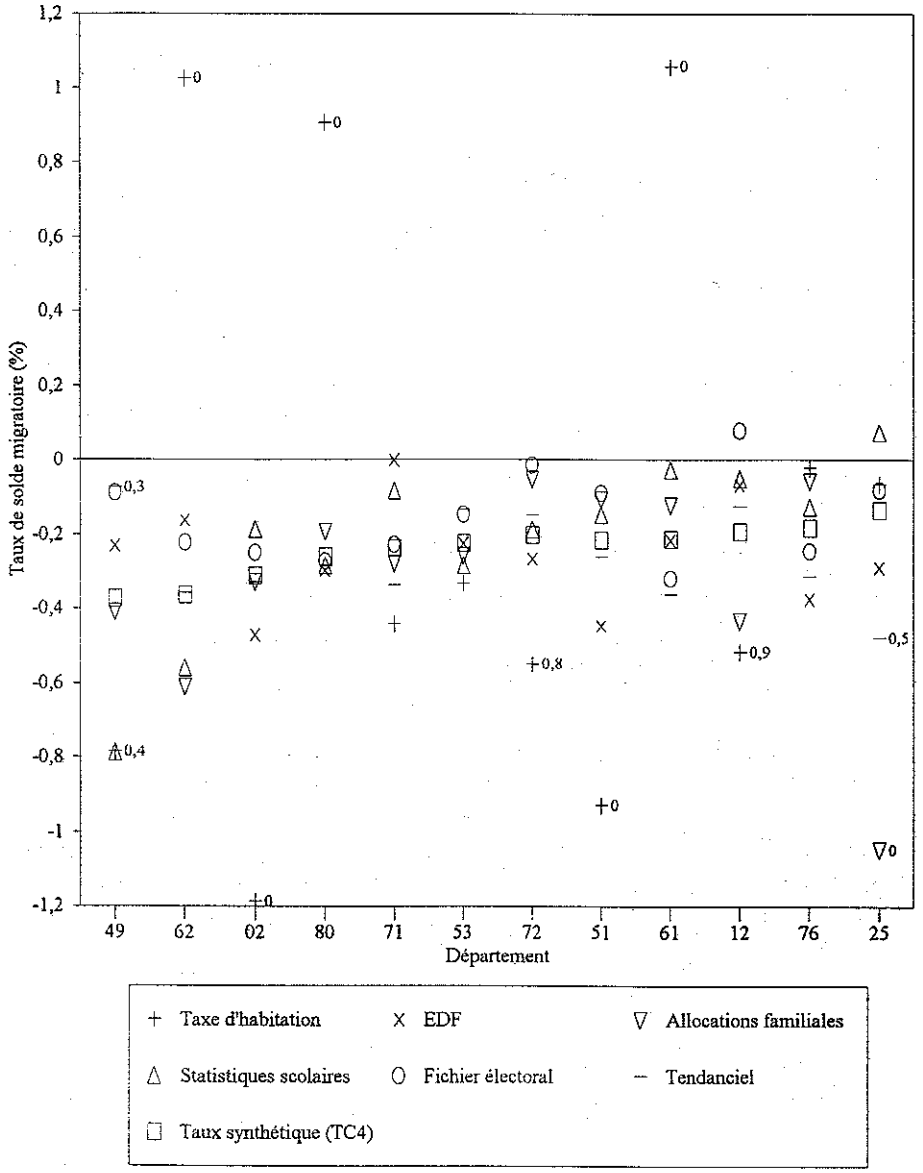

Figure 1 : Synthèse des taux de solde migratoire de l'année 1990 pour douze départements, repérés par leur numéro (49, 62...).

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.

Tableau 7
Moyenne des écarts dans le test rétrospectif

Ensemble des taux

	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

Notes : - 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 été perturbée par la procédure «IR-TH» à partir de 1987.
- Les valeurs des écarts correspondent à des taux exprimés en %.

Cela n'a d'ailleurs rien d'étonnant puisque la plupart des méthodes ont été sensiblement améliorées par rapport au test. 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.

Cependant, avant de passer à la production routinière d'estimations, une phase de mise au point et d'adaptation sera nécessaire.

9.1. Détermination des poids et des normes.

La détermination des poids et des normes est un point central du système. Notons d'ailleurs que seuls importent les poids relatifs des différentes sources. 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.

Pour les sources testées sur la période intercensitaire 1982-1990, les poids retenus pour initialiser le système «post-1990» au niveau départemental reflètent la précision à moyen terme de chaque source : ils sont inversement proportionnels à la

variance des erreurs commises en 1990 (après correction des anomalies annuelles). Ce mode de détermination ne tient pas compte de la non-indépendance des taux issus de certaines sources. En général, les corrélations entre les erreurs commises en 1990 avec les différentes sources sont faibles. Les deux méthodes TH et «abonnés électriques», qui font intervenir la même estimation de la taille moyenne des ménages (TMM), constituent une exception ; le coefficient de corrélation est voisin de 0,6. Une question se pose donc : ne faut-il pas faire en sorte que la somme des poids appliqués à ces deux sources ne dépasse jamais une certaine limite, très sensiblement inférieure à la somme des poids «a priori» ? La réponse est sans doute positive ; avec un coefficient de corrélation voisin de 0,6, la limite en question devrait même correspondre à un abattement d'environ 40 %. Toutefois, l'estimation de la TMM ayant été améliorée par rapport au test rétrospectif, la corrélation des erreurs devrait être plus faible. Là encore, il faut attendre le prochain recensement pour le savoir. D'ici là, l'analyse des résultats obtenus pour les premières années de la période 1990-1999 devrait fournir des indications utiles. Pour l'année 1990, il n'y a pas de corrélation positive entre les écarts présentés par les deux sources. Mais il serait hasardeux de ne considérer qu'une seule année ; d'autant que la source TH présente beaucoup d'anomalies (éliminées de la corrélation) en 1990.

Le choix des normes départementales retenues pour l'initialisation du système repose à la fois sur les écarts (entre taux de solde migratoire issus de chaque source et taux synthétique) observés dans le test et sur une relation supposée de quasi-proportionnalité entre le poids et l'inverse du carré de la norme. Dans l'ensemble, les écarts constatés sur les taux de l'année 1990 semblent assez cohérents avec les normes retenues. Cependant, pour certaines sources, la source «fichier électoral» notamment, ces écarts sont sensiblement inférieurs aux normes. Il y aura donc sans doute lieu de réviser les normes et les poids. Mais il est préférable de le faire en se fondant sur les résultats de plusieurs années.

Pour une source nouvelle, 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. On peut alors adapter la norme en fonction des écarts obtenus et adapter le poids en conséquence, en se servant, faute de mieux, de la relation de quasi-proportionnalité entre le poids et l'inverse du carré de la norme, déjà mentionnée.

Au niveau départemental, il ne semble pas utile d'adapter les 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. Les analyses semblent montrer qu'une fonction du type suivant peut 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. Pour ce faire, il est suggéré d'utiliser une méthode itérative analogue à celle indiquée plus haut, en se servant encore de la même relation entre poids et normes.

Une fois les poids et les normes adaptés ou définis comme il vient d'être suggéré, il est recommandé de les conserver jusqu'au prochain recensement ; à moins que les analyses annuelles ne montrent une évolution marquée, pour telle ou telle source, de l'indicateur de distance sur lequel repose la détermination de la norme.

9.2. Traitement de certaines situations particulières.

9.2.1. Difficulté de convergence.

Le processus de détermination du taux synthétique converge en général assez rapidement. Les tests menés au niveau départemental sur 1982-1990 ont montré que les taux étaient très souvent stabilisés à partir de la quatrième itération. L'essai réalisé pour l'année 1990 le confirme. Il arrive cependant, dans certaines situations, que la convergence soit difficile. Dans les quelques cas rencontrés, la poursuite des itérations finit par aboutir à un résultat stable, mais pas nécessairement acceptable. On peut toujours poursuivre les itérations. Il semble toutefois judicieux, en cas de convergence difficile, de provoquer un signal d'alerte, d'examiner la situation et, le cas échéant, d'intervenir ponctuellement.

9.2.2. Annulation de tous les poids.

Une situation de blocage peut être créée par l'annulation de tous les poids, y compris celui de l'estimation tendancielle. Là encore, il semble judicieux de provoquer un signal d'alerte. Toutefois, il faut aussi prévoir un dispositif automatique pour éviter le blocage. Une solution, simple et toujours applicable, consiste à prendre le taux tendanciel comme taux synthétique, lorsque, la somme des poids étant nulle, ce dernier ne peut être calculé.

9.2.3. Intervention ponctuelle.

Dans les deux situations précédentes une intervention ponctuelle peut être utile, voire indispensable. Il peut d'ailleurs en être de même dans d'autres cas où le taux

synthétique final issu du processus automatisé semblerait discutable. On propose d'introduire cette possibilité de la façon suivante, qui est rationnelle, sans risque et qui s'intègre bien au système : faire intervenir, pour chaque source, un coefficient de modulation supplémentaire, qui vaudrait 1 par défaut, mais qui pourrait être diminué, voire annulé, à la discrétion du gestionnaire, en cas de nécessité.

10. 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.

Bibliographie

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.

DEKNEUDT, J. (1990), Migrations à l'âge scolaire et évaluations de population, Département de la démographie, note interne n° 13/F127, INSEE, Paris.

FONTAINE, F. (1986), « Estimer la population d'une région à partir de l'emploi et du chômage : l'exemple du Nord-Pas-de-Calais », *Economie et statistique*, n° 193-194, INSEE, Paris.

GUEGUEN, 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*. John Wiley, New-York.

LAURENT, L., et GUEGUEN, 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.

MEURIC, L. (1995), Précision des estimations locales de population fondées sur le nombre de personnes par ménage tiré des enquêtes annuelles sur l'emploi, Division Emploi, note interne n° 214/F232, INSEE, Paris.

SAUTORY, O. (1993), La macro CALMAR - redressement d'un échantillon par calage sur marges, Direction des statistiques démographiques et sociales, document de travail n° F 9310, INSEE, Paris.

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

STATISTIQUE CANADA (1995), Rapport sur la méthodologie de production des données migratoires à partir des dossiers d'impôt, Division des données régionales et administratives, Ottawa.