Bernard AUBRY
INSEE DR Alsace

P R O J E T

A propos de la qualité du recensement de 1999

(version de juin 2008)

Cette note se propose de mettre en évidence un certain nombre d'anomalies observées dans les fichiers détails du recensement de 1999. Le contenu de cette note concerne plus particulièrement les résultats d'une exploitation de fusion des deux fichiers d'exploitation utlisés par les chargés d'étude de l'Insee et servant à la diffusion des données (notamment sur EDL).
Le problème n'est pas nouveau, il s'est révélé à chaque recensement. On sait en effet que les données issues de l'exploitation par sondage (généralement au 1/4) diffèrent de celles issues des exploitations complémentaires, mai sà ma connaissance la mesure précises des divergences n'a jamais été faite, ou du moins si elle l'a été, elle n'a jamais fait l'objet d'une diffusion. .

Si les erreurs ont toujours existé, leurs conséquences seraient aujourd'hui plus graves dans la mesure où les utilisateurs des chiffres sont plus exigeants parce que mieux formés. Ces erreurs troublent les utilisateurs de chiffres, qui doivent de surcroit être attentifs à éviter de tomber dans l'un ou l'autre des multiples pièges liées à l'interprétation des statistiques (l'effet de structure par exemple).

Il arrive en effet que l'on diffuse, à partir d'une même source, à une même date, pour un même lieu, des chiffres contradictoires.

Pour prendre le cas de la commune de Strasbourg, l'écart, d'après l'EDL, entre le nombre d'emplois mesurés dans l'exploitation complémentaire est **inférieur de 2 50**1 à ce qu'il est dans l'exploitation principale (- 1,7%). Quand il s'agit des seules navettes sortantes, l'écart est **supérieur de 2 142** (+ 8,1% par rapport à l'exploitation principale)

Le problème n'est pas nouveau. Je ne pense pas être le seul chargé d'études à avoir entendu les doléances de ceux à qui les données recensements sont destinées. J'ai été maintes fois confronté aux récriminations de notre « clientèle ». Dans une note datée de décembre 1986 retrouvée récemment, j'avais écrit : « *Aux recensements de 1975 et 1982, les différences enregistrées selon les modes de tirage ont parfois été si fortes qu'elles interdisaient presque toute analyse un tant soit peu approfondie des résultats* ». C'était il y a 30 ans, soit 13 ans avant que l'on engage à la DR d'Alsace un travail d'harmonisation des recensements qui ont eu lieu depuis 1962.

Nous avions alors en effet beaucoup de difficultés à comprendre le fonctionnement du marché du travail des zones frontalières de Wissembourg et de Saint-Louis, deux zones d'emploi qui fournissent à l'Allemagne et à la Suisse des effectifs de frontaliers nombreux et, à l'époque, en très forte croissance. Cette main d'œuvre était puisée non seulement sur les ressources propres des territoires, mais également sur les régions voisines et même plus lointaines. La confrontation des chiffres issus des exploitations par sondage et exhaustive semait la plus grande confusion. En revanche une perspective historique, basée uniquement sur les chiffres des sondages permettait dans une certaine mesure de gommer les aspérités dues aux incertitudes aléatoires. Le besoin d'y voir clair dans des zones assez complexes a joué un rôle décisif dans la décision de créer le fichier Saphir[1] ?

Il reste qu'il est toujours un peu difficile de trouver l'écho souhaité quand on cherche à faire passer un message pour une plus grande transparence dans la présentation des données diffusées. On ne semble pas, me semble-t-il, très sensible à ces problèmes de cohérence interne des données. Est-ce parce que les statisticiens nationaux sont habitués à travailler sur des territoires très peuplés qu'ils considèrent comme négligeables les écarts entre données issues des traitements exhaustifs et par sondage ? Est-ce parce que le souci de fournir une information locale de qualité n'était pas alors aussi marqué qu'il l'est aujourd'hui ?

Dans le but de mettre en évidence les divergences de codification, un fichier spécifique avait été créé à la DR d'Alsace[2]. Il consistait à fusionner une partie des deux fichiers-détails disponibles, celui de l'exploitation principale et celui de l'exploitation complémentaire. C'était en 2002. Comme à l'époque on m'a demandé de ne pas diffuser les résultats issus de cette confrontation, je n'ai pas été plus loin dans l'investigation. Je l'ai reprise récemment, à la faveur de l'entrée en vigueur du *« Code de bonnes pratiques de la statistique européenne»*.

C'est l'objet de cette note de diffuser certains résultats de cette exploitation. Elle devrait intéresser mes collègues régionaux qui travaillent sur le thème de l'analyse territoriale. Dans le domaine du marché du travail (population active et emploi), on trouvera notamment la liste des communes à problème.

---

[1] *Saphir : Système d'analyse de la population par l'historique des recensements.*

[2] *Référence du fichier, sur le* ***CNIO : SP72.L00.SAP3.RP99.FUSION,*** *table AA (57 226 208 enregistrements)*

Les données des enquêtes annuelles des recensements sont peu à peu utilisables au plan local, du moins pour les agents de l'Insee. Quand tomberont les chiffres de l'emploi par commune, *à quels chiffres du RP99 pourrons-nous les raccrocher ?*. Selon qu'ils seront comparés aux données de l'exploitation principale (l'exhaustif) ou aux données de l'exploitation complémentaire (le sondage), les conclusions sur l'évolution du marché de l'emploi seront, pour un grand nombre de communes tout au moins, radicalement différentes.

Les différentes sources d'erreurs, les différents pièges

Avant de présenter ces résultats, il est utile de rappeler les principaux types d'anomalies que l'on rencontre dans un recensement de la population :

° des défauts de couverture, du fait que certains individus qui font partie du champ ont échappé à la collecte, ou bien sont comptés deux fois. Ces critiques, bien souvent évoquées, y compris parfois dans les médias, ne sont pas abordées dans ce document.

° des défauts de collecte (erreurs dans les déclarations) et de codification. Ainsi lorsque tous les membres d'un même établissement reçoivent une même date de naissance ou une même localité de résidence antérieure. Bien que ce ne soit pas l'objet premier de cette note, on donne quelques exemples d'anomalies qu'il aurait été facile d'éviter par un traitement statistique ad hoc.

° enfin des divergences de codification **internes** à une même enquête. On sait en effet que les bulletins du recensement sont exploités en plusieurs phases, dont deux conduisent à des tableaux de résultats : l'exploitation principale (on dit aussi exploitation exhaustive ou légère car elle ne porte que sur une partie des variables) et l'exploitation complémentaire (dite aussi exploitation lourde ou par sondage effectuée par tirage de logements).
Certains logements (les « trois-quarts ») ne sont donc codifiés qu'une seule fois, les autres (le « quart ») le sont deux fois. Certaines variables sont, par construction, codifées de la même façon dans les deux exploitations (le sexe, l'âge, le diplôme, etc). Pour d'autres, il n'en est pas de même. Les affectations attribuées aux individus dans l'une ou l'autre modalité d'une même nomenclature sont susceptibles de diverger. C'est le cas notamment pour la population active (code « emploi » et « commune de travail »), c'est le cas aussi pour la résidence antérieure.

° A ces anomalies s'ajoutent les incertitudes aléatoires, liées au fait que le traitement complémentaire se fait par tirage d'un échantillon de logements.

° On pourrait aussi évoquer des problèmes liés à des défauts d'anticipation. Je n'ai pas capacité d'en parler ici de façon approfondie. Néanmoins, on pense par exemple aux conséquences liées aux changements de nomenclature d'activité économique (passage de la NAP à la NAF en 1993, soit au cours de la période intercensitaire), ce qui a rendu périlleuse toute publication d'évolution de l'emploi à un niveau de détail géographique ou structurel un peu fin. Or l'évolution de l'emploi est incontestablement l'une des préoccupations premières des responsables territoriaux.

Principaux résultats

La population active et l'emploi

Il s'agit là du domaine où les anomalies sont les plus fréquentes et les plus préjudiciables à une appréciation satisfaisante du marché du travail local. L'affectation du lieu de travail pour les actifs se fait de façon différente dans les deux traitements. Dans l'exploitation exhaustive, on reprend directement l'indication de la commune figurant sur le questionnaire. Dans l'exploitation par sondage, l'affectation du lieu de travail se fait automatiquement après identification de l'établissement où travaille la personne. Pour prendre un exemple extrême, la plupart de salariés de la principale usine Peugeot d'Alsace ont indiqué Mulhouse comme lieu de travail alors qu'en toute rigueur l'implantation de l'établissement se situe dans une commune limitrophe (Sausheim). *Ces divergences pourraient être sans grande importance si les deux communes apparaissaient réunies dans tous les regroupements territoriaux dans lesquels la statistique est convoquée. Cela n'est pas le cas.*

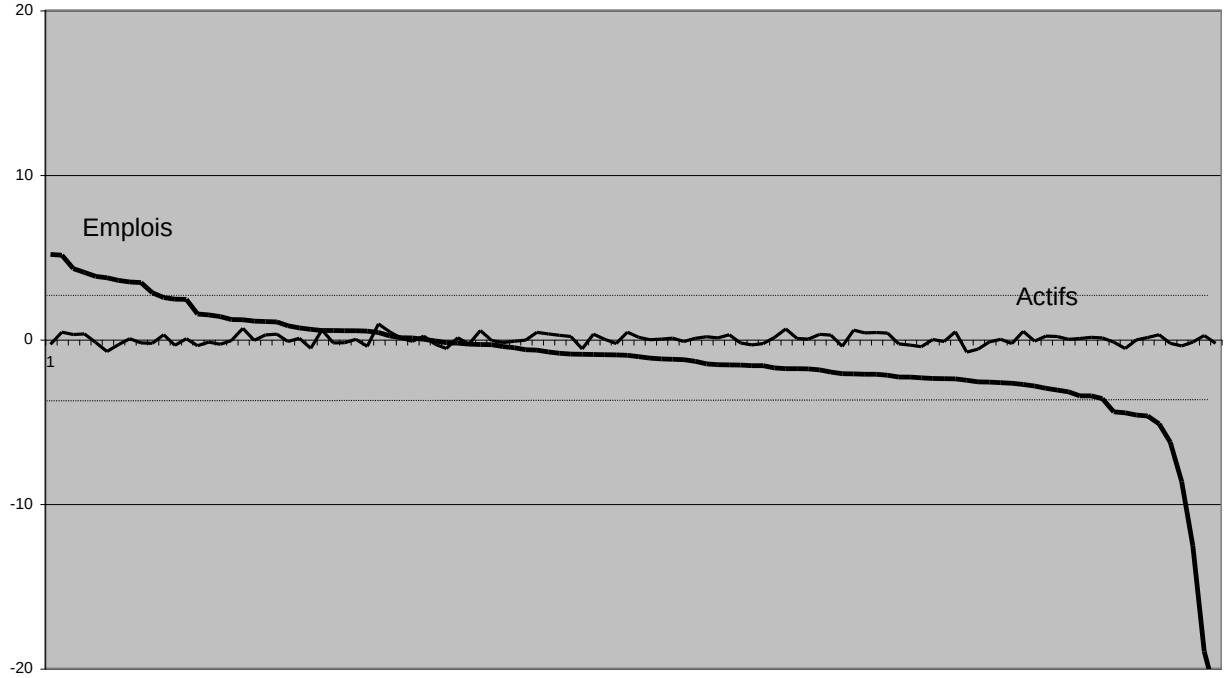
On peut penser que la qualité de la codification du sondage est meilleure que celle de l'exhaustif. Elle l'est en effet. Mais il existe des cas où c'est l'inverse qui se passe. Ainsi pour certains établissements militaires. Or l'affectation de tout un établissement de l'armée dans l'une ou l'autre commune, si elle se fait différemment à deux recensements différents conduit le chargé d'étude à des interprétations différentes de l'évolution du marché du travail du territoire d'étude. Les conséquences sont évidemment sensibles au niveau de la commune. Elles le sont même au niveau d'une zone d'emploi.

L'analyse proposée porte sur les communes, sachant qu'elle peut être appliquée également à des territoires plus vastes (zones d'emploi par exemple). On distingue deux points de vue : la population active comptée au lieu de résidence et la population active comptée au lieu de travail (l'emploi). Les deux ensembles comprennent une partie commune (la population qui réside et travaille dans la localité d'étude) et une partie complémentaire (dans un cas les navettes sortantes et dans l'autre les navettes entrantes).

On observe que les divergences sur la population active sont généralement assez faibles, ce qui est la conséquence du mode de tirage de l'échantillon. En revanche ils sont parfois très importants sur l'emploi comme le montre les graphiques qui suivent et qui portent sur les communes de plus de 50000 habitants au RP99.

Graphique I - l'emploi et la population active (écarts normés)

On détermine la différence entre les effectifs de l'exploitation complémentaire et ceux de l'exploitation principale pour la population active d'une part et l'emploi d'autre part. Les écarts sont « normés » pour tenir compte de la différence de taille entre les communes. Pour calculer ces écarts normés, on prend pour écart type la formule figurant dans les fascicules verts de diffusion des résultats de l'exploitation complémentaire. En principe tout écart normé supérieur, en valeur absolue, à 3 suggère l'existence d'une anomalie. Les communes sont classées selon la valeur décroissante de cet écart normé.

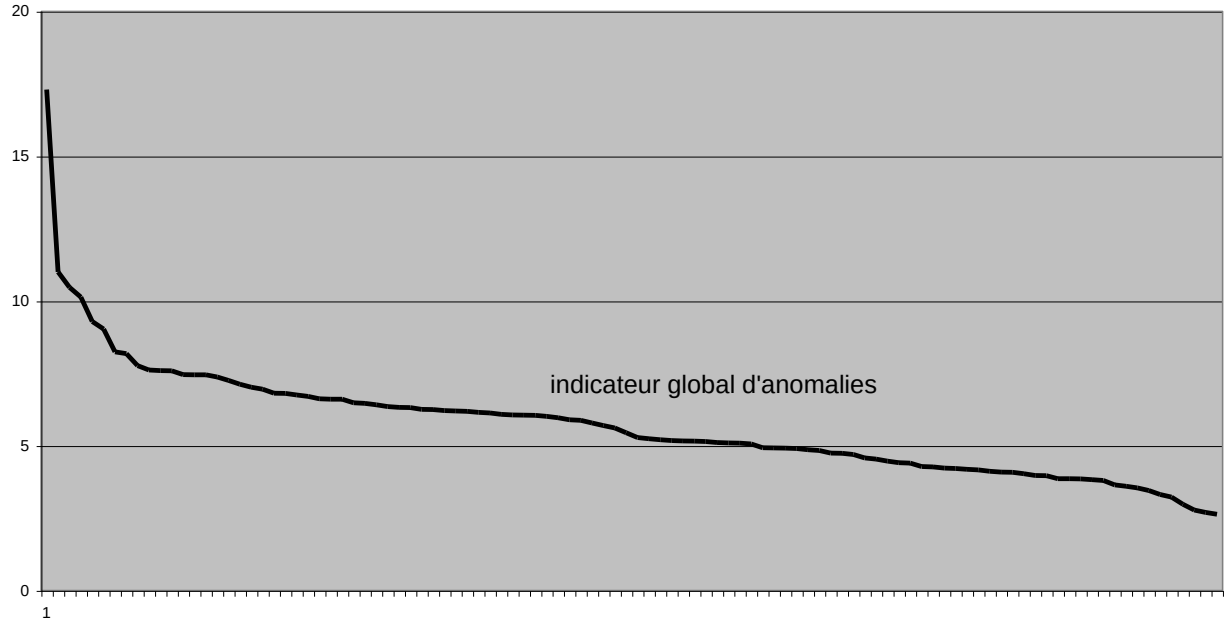


On constate que les écarts normés entre les deux traitements sont toujours faibles, voire insignifiants lorsqu'il s'agit de la population active résidente (cela est dû au mode de tirage équilibré des logements). En revanche il est parfois très élevé pour l'emploi. *Dans un cas comme dans l'autre la formule de référence pour l'écart type de la loi normale ne s'applique pas.*

Cet indicateur est une résultante. Il se peut que cet indicateur soit faible, voire quasiment nul, non pas parce qu'il n'y aurait pas d'anomalies, mais parce que les erreurs se compensent. C'est pourquoi on présente également des taux globaux d'anomalies par commune.

Graphiques II - Le taux global d'anomalies (en %)

Pour chaque commune on détermine les anomalies pour les trois composantes de la population concernée : les personnes qui résident et travaillent dans la commune, les entrants et les sortants. Ces trois sous-ensembles sont formés d'éléments distincs. Le taux d'anomalies est calculé en divisant la somme des erreurs par l'effectif total des trois groupes.
Tous ces résultats sont détaillés dans l'annexe où l'on prend l'exemple de la commune de Strasbourg.



Les communes ayant les valeurs extrêmes figurent sur le tableau suivant :

communes	**Taux normé (pour l'emploi)**	**Taux global d'anomalies %**
Annecy	5,2	9,3
Champigny/M	5,1	9,0
…	…	…
Cergy-Pontoise	3,7	11,0
…	…	…
Besançon	- 5,1	4,5
Lyon	- 6,2	6,2
Aulnay-sous-Bois	- 8,6	10,1
Courbevoie	- 12,6	17,3
Mulhouse	- 19,0	10,5
Paris	- 21,6	4,9
…	…	…
Marseille	- 4,5	2,2

Les communes ayant les écarts normés sur l'emploi les plus élevés ne sont pas nécessairement les mêmes que celles qui ont les taux globaux d'anomalies les plus élevées, car ce n'est pas le même concept.

Le code « emploi » (EMPL)

S'agissant de la classification des emplois en 10 postes, on rencontre des divergences parfois assez fortes sur le total des effectifs d'une même modalité. Prenons l'exemple des travailleurs indépendants. Ils sont 382 000 à l'exhaustif et 372 600 dans le sondage. L'écart pourrait paraître faible. Il l'est en effet si l'on se limite à l'examen des effectifs totaux. En réalité quand on regarde davantage dans le détail, on constate que l'écart d'environ 10 000 personnes est la résultante de deux divergences qui se compensent. On trouve à l'exhaustif 44 000 personnes dans ce groupe qui ne sont pas codifiées telles au sondage et, inversement, 35000 personnes sont codifiées dans le sondage comme tels mesurés dans l'exhaustif. Il va de soi que rien ne prouve que ces anomalies se répartissent de façon régulière dans les différentes catégories de population (sexe, âge diplôme…). Des analyses plus approfondies de ce fichier permettraient de mieux caractériser ces divergences.

Les frontaliers

La demande locale en matière d'information sur les frontaliers est très grande et le sujet est sensible. On est tenté de considérer comme travailleurs frontaliers les personnes qui désignent un lieu de travail hors de la métropole. Or d'un côté on dénombre environ 280 000 personnes, de l'autre près de 250 000. Le traitement par sondage a donc réduit de 30 000 à peu près le nombre de ces actifs travaillant hors de la métropole. Là encore, certaines catégories et certaines zones géographiques sont plus touchées que d'autres par les divergences dans les résultats.

La résidence antérieure

La localisation de la résidence antérieure de l'enfant né entre deux recensements n'étant pas cohérente dans les deux exploitations (dans l'une on prend par convention le lieu de résidence antérieure de la personne de référence du ménage, dans l'autre cas celle de la personne de référence de la famille), il en résulte des divergences fréquentes. Pour l'ensemble des enfants de 0 à 9 ans, ces incohérences touchent environ le quart des effectifs.

Questions et suggestions.

a) Faut-il continuer à diffuser simultanément les résultats des deux traitements, au risque d'embrouiller les utilisateurs dans leurs travaux d'analyse ?

Si oui, ne faudrait-il pas prévoir la rédaction d'une note détaillée présentant le problème et listant les principaux cas d'anomalie ?.

b) S'agissant plus particulièrement des données sur la population active et l'emploi :

Est-il possible de corriger les données, pour créer une nouvelle base qui serait affranchie des principales incohérences ?

A priori, ce devrait être possible. Il faudrait alors, à partir d'une analyse statistique des cas anormaux, confronter les données des deux exploitations du RP99 avec celles du fichier des DADS de l'année correspondante. Ayant repéré les établissements concernés, il devrait être assez facile d'établir un algorithme de correction automatique des lieux de travail. De toute façon les résultats devraient être soumis aux régionaux pour examen.

La même opération pourrait être faite pour les recensements antérieurs au RP99.

Une réflexion supplémentaire, inspirée d'un article lu récemment quelque par : ***« Dans le monde actuel la qualité n'existe qu'à la condition de fournir des preuves de sa mise en œuvre et d'accepter les regards externes ».***

Extrait du code des bonnes pratiques de la statistique européenne

Principe 4 - Engagement sur la qualité

La qualité des produits est régulièrement contrôlée selon les critères du SSE. Des procédures sont prévues pour assurer le suivi de la qualité de la collecte, du traitement et de la diffusion des statistiques. Les principales productions statistiques font l'objet d'une évaluation régulière et approfondie, le cas échéant, en faisant appel à des experts extérieurs.

Principe 6 - Impartialité et objectivité

Les erreurs découvertes dans des statistiques déjà publiées sont corrigées dans les meilleurs délais.

Principe 7 : Méthodologie solide

Des statistiques de qualité sont fondées sur une méthodologie solide. Cela nécessite des procédures, des compétences et des outils adéquats.

Principe 8 : Procédures statistiques adaptées.

Des systèmes informatiques appropriés sont utilisés pour l'imputation et l'apurement ; ils sont régulièrement évalués, corrigées ou mis à jour le cas échéant.

Principe 10 - Rapport coût - efficacité

Les opérations de routine (par exemple la saisie, a codification ou la validation) sont

automatisées dans la mesure du possible. Les possibilités offertes par les technologies de l'information et de la communication sont exploitées de façon optimale dans la collecte, le traitement et la diffusion de données

Principe 12 : Exactitude et fiabilité

Les données collectées, les résultats intermédiaires et les productions statistiques sont évalués et validés. Les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage sont analysées et systématiquement documentées conformément aux différents critères de qualité du SSE. Les révisions font systématiquement l'objet d'études et d'analyses, qui sont utilisées en interne pour alimenter les processus statistiques

Principe 14 - Cohérence et comparabilité

Les statiques présentent une cohérence interne (par exemple, vérifiant les égalités arithmétiques et comptables. Les statistiques sont cohérentes et peuvent être rapprochées sur une durée raisonnable.

Principe 15 : Accessibilité et clarté

Les statistiques sont présentées sous une forme qui facilite une interprétation correcte et des comparaisons utiles. Les utilisateurs sont tenus informés des aspects méthodologiques relatifs aux procédures statistiques et de la qualité des résultats statistiques par rapport aux critères de qualité du SSE.

A propos de la qualité du recensement de 1999 (version juin 2008)

Annexe

Plan

I - De la population active à l'emploi

Compte tenu de l'importance prise par les déplacements domicile-travail, on ne peut pas se contenter de repérer les anomalies relatives à la seule population résidant sur un territoire. Il faut aussi repérer les anomalies de la population active au lieu de travail, c'est-à-dire à l'emploi. Or, dans le débat social, la question de l'emploi et de son évolution est d'importance. Elle l'est tout particulièrement au plan local, et c'est en matière d'emploi que les divergences sur les effectifs entre les deux traitements sont précisément les plus fortes.

La présentation des anomalies s'avère donc relativement compliquée. S'agissant d'un territoire pris comme référence (on pourra prendre aussi bien une commune, une zone d'emploi ou n'importe quel zonage), on doit faire intervenir trois catégories d'actifs :

° les personnes qui résident et travaillent dans le territoire (R & T)
° les personnes qui sortent du territoire pour aller travailler ailleurs (navettes sortantes)
° les personnes qui résident hors du territoire et viennent y travailler (navettes entrantes)

Après avoir décomposé les données globales, on présente quelques indicateurs susceptibles de caractériser la fiabilité des chiffres.
L'exemple chiffré porte sur la commune de Strasbourg (commune de référence).

Remarques

La compréhension des pages suivantes sera grandement facilitée si l'on se reporte à la partie III qui décrit la méthode. Notons toutefois d'ores et déjà que le fichier de fusion sur lequel portent les résultats comprend deux blocs :

- Le bloc dit des 3/4 qui correspond à la population traitée uniquement de façon exhaustive. Il ne peut y avoir par construction d'incohérences par construction.
- Le bloc dit du 1/4 qui comprend les ménages et les individus dont les informations ont fait l'objet d'une double codification : à l'exhaustif d'abord, puis au quart quelques mois plus tard[3].
  Pour chaque individu présent dans le 1/4, on dispose de deux pondérations. L'une pour l'exploitation principale (toujours égale à 1), l'autre pour l'exploitation complémentaire (en général 4, parfois 1).

D'une façon générale, tous les symboles se terminant par 1 concernent les données traitées dans le cadre de l'exploitation principale (par exemple DCLT1), ceux se terminant par 4 correspondent à des données traitées dans le cadre de l'exploitaiton complémentaire (ex : DCLT4).

[3] *En toute rigueur, on ne devrait pas parler d'exploitation exhaustive (pour le traitement principal) et d'exploitation par sondage au 1/4 (traitement complémentaire) puisque dans certaines cas, tout ou partie d'une même commune a fait l'objet d'un traitement exhaustif dans la phase complémentaire. Mais ces cas sont relativement rares. Pour justifier cet emploi de termes impropres, on fera remarquer qu'ils ont l'avantage d'être bien compris, notamment par les « anciens » qui ont pratiqué les recensements antérieurs. C'est pourquoi, pour alléger l'écriture, on acceptera l'emploi quelque peu abusif de termes exhaustif et sondage.*

1) Résider et travailler dans la commune.

Le tableau que l'on propose pour mettre en évidence les anomalies se compose de deux parties :

☞ la partie supérieure (cohérence) correspond au cas où les modalités du lieu de travail sont identiques dans les deux traitements (s'agissant des 3/4, la question de la cohérence ne se pose même pas par construction).

☞ la partie inférieure correspond aux situations d'incohérence.
En colonne, on distingue les résultats obtenus selon l'un ou l'autre des traitements (exhaustif ou par sondage).
S'il s'agit du sondage, il convient de distinguer les effectifs bruts (nombre d'individus rencontrés) et les effectifs redressés (le coefficient de pondération est en gros de 4)

Tableau 1- Résider et travailler à Strasbourg.

Qualité des données	Codes QQ - QT	origine	Cumul exhaustif	Cumul sondage*	
				redressé	brut
cohérence	0 - 19	3 / 4	**54 363**	-	
	1 - 14	1 / 4	**19 819**	**71 653**	19 819
			74 182	**71 653**	19 819
incohérence	9 - 10	1 / 4	**818**	2 885	818
	9 - 04	1 / 4	334	**1 2**04	334
Totaux			**75 000**	**72 857**	20 123

Les chiffres en gras correspondent aux effectifs pris en compte dans les totalisations

Mode de lecture

☞ *Effectif total, selon <u>l'exhaustif</u> : **75 000** personnes.*
Le chiffre est obtenu par addition de trois éléments :
- *les personnes appartenant aux trois quarts (54 363, par construction hors champ du sondage) correspondant à QT=19 (voir annexe 3).*
- *Les personnes du quart (19 819), codifiées de façon cohérente dans les deux exploitations ; elles correspondent à QT=14.*
- *Les personnes codifiées de façon incohérente (818). Dans le fichier issu du sondage, elles ont été considérées comme ne travaillant pas à Strasbourg alors que dans le fichier exhaustif, elles avaient pour lieu de travail Strasbourg (QT=10).*

☞ *Effectif total, selon le <u>sondage</u> : **72 857** personnes.*
Le chiffre est la somme de deux éléments :
- *les personnes codifiées correctement dans les deux exploitations : 71 653 (QT=14)*
- *les personnes codifiées de façon incohérente : 1 204 personnes (elles ont pour lieu de travail Strasbourg dans le sondage et hors Strasbourg dans l'exhaustif (QT=04). Notons que 1 204 est un chiffre redressé, puisque le nombre d'enregistrements décomptés est de 334 (colonne cumul exhaustif).*

☞ ***On remarque que l'effectif du sondage est très inférieur à celui de l'exhaustif : <u>2143 personnes d'écart.</u>***

En se limitant à la partie sondages au 1/4, on calcule les valeurs suivantes :

Incohérences e1 = 818 et e4 = 1204
Poulation de référence : rf1 = 20 637 (19 819 + 818)
et rf4 = 72 857 (71653 + 1 204)

d'où le calcule de deux taux : tx1= e1/rf1 et tx2 = e4/rf4

Tableau 1a- des résultats comparés pour quelques communes

commune		**Incohérence**		**Effect. de référence**		**Taux d'erreurs (%)**	
		E1	**E4**	**RF1**	**RF4**	**Tx1**	**Tx4**
67482	Strasbourg	818	1 204	20 637	72 857	4,0	1,7
68224	Mulhouse	669	523	6 895	22 928	**9,7**	2,3
74011	Annecy le V	505	164	1 193	2 916	**42,3**	5,6
75056	Paris	1 010	18 226	175 327	687 420	0,6	2,7
92026	Courbevoie	133	1 476	2 048	9 136	6,5	**16,2**

Bien noter que ce tableau ne prend en compte que des individus du bloc 1/4. On remarque que tant la population de référence (RF4) que les erreurs (E4) sont grosso modo 4 fois plus élevés que les deux autres (RF1 et E1). On retrouve pour Strasbourg les effectifs figurant dans le tableau 1.

2) Sortir de la commune

Le groupe des « sortants » (population résidant à Strasbourg et n'y travaillant pas) peut être traité de la même façon.

Tableau 2 - Sortir de Strasbourg :
(population résidant à Strasbourg et n'y travaillant pas)

Qualité des données	QQ- QT	origine	Cumul exhaustif	Cumul sondage redressé	Cumul sondage brut
cohérence	0 - 99	3 / 4	**18 887**	-	-
	1 - 00	1 / 4	**6 766**	**24 763**	
			25 653	**24 763**	
incohérence	9 - 04	1 / 4	**334**	1 204	334
	9 - 00	1 / 4	**181**	**661**	181
	9 - 10	1 / 4	818	**2 885**	818
Totaux			**26 168**	**28 309**	7 765

Le tableau a presque la même structure que le précédent. Toutefois dans la partie incohérence, on doit faire apparaître trois cas de figure au lieu de deux, à savoir :

- QT = 04 : les actifs sortent de Strasbourg selon l'exhaustif, mais non selon le 1/4
- QT = 10 : l'inverse de QT=04, à savoir ces personnes sortent de Strasbourg selon le quart, mais non selon l'exhaustif.
- QT = 00 : dans les deux exploitations il y a sortie de Strasbourg, mais vers des communes différentes.

☞ ***La différence sur le total est encore très importante : <u>2141 personnes.</u>***

Tableau 2a- Des résultats pour quelques communes

communes		incohérences		Eff. de référence		Taux d'erreurs	
		e1	e4	rf1	rf4	ttx1	tx4
67482	Strasbourg	515	3 546	7 281	28 309	7,1	**12,5**
68224	Mulhouse	343	3 054	4 568	18 277	7,5	**16,7**
75056	Paris	9 302	21 465	78 476	287 697	**11,9**	7,5
74011	Annecy le V.	91	2 220	853	5 268	**10,7**	**42,1**

On remarque pour Strasbourg un taux d'erreurs élevé pour tx4 : cela signifie que 12,5% des personnes qui dans le sondage quittent Strasbourg n'en sortent pas dans l'exhaustif. L'anomalie est encore plus flagrante pour Annecy-le Vieux puisque le taux d'anomalies est de 42,1*%. S'agissant de la commune de Paris, les taux d'erreurs sont relativement importants surtout du côté de tx1.*

3) Entrer dans la commune

Tableau 3 - Entrer dans Strasbourg

(personnes travaillant à Strasbourg mais n'y résidant pas)

Qualité des données	QQ - QT	origine	Cumul exhaustif	Cumul sondage	
				redressé	brut
cohérence	0 - 99	3/4	**52 358**		
	1 - 00	1/4	**21 654**	**71 286**	21 654
			74 012	**71 286**	21 654
incohérence	9 - 04	1 / 4	**728**	2 423	728
	9 - 00	1 / 4	**336**	1 134	336
	9 - 10	1 / 4	645	**2 091**	645
	9 - 00	1 / 4	407	**1 370**	407
Totaux			**75 076**	**74 747**	22 718

La lecture du tableau se fait de façon analogue aux précédents. On remarque que cette fois il y a quatre lignes pour caractériser les incohérences :

- *QT = 04 : entrées à Strasbourg à l'exhaustif, mais non au 1/4, puisque DCR = DCLT4, ce qui signifie l'absence de navette.*
- *QT = 10 : entrées à Strasbourg au 1/4, mais non à l'exhaustif (avec DCR = DCLT1)*

- *QT = 00 : changement de commune, mais Strasbourg est le lieu de travail à l'exhaustif (pour le quart, DCR ne DCLT4 (la personne change de commune, mais ne travaille pas à Strasbourg).*
- *QT = 00 : changement de commune, mais Strasbourg est le lieu de travail au sondage.*

☞ ***la différence entre les effectifs est de <u>329 personnes</u> (chiffres du sondage inférieurs à ceux l'exhaustif.***

Tableau 3a- des résultats pour quelques communes

communes		incohérences		effectifs de référence		Proportion d'erreurs	
		e1	e4	rf1	rf4	tx1	tx4
67482	Strasbourg	1 064	3 461	22 718	74 747	4,7	4,6
68224	Mulhouse	2 388	834	11 572	30 112	**20,6**	2,8
68300	Sausheim	55	9 255	1 133	12 742	4,9	**72,6**
75056	Paris	24 208	18 795	258 343	895 176	9,4	2,1
91182	Courcouronnes	49	3 776	577	2 795	8,5	**65,2**

Le tableau fait apparaître deux cas extrêmes. Celui de Sausheim (usine Peugeot) , près de Mulhouse (tx4 = 72,6 %) : sur un total de 12 742 personnes entrant à Sausheim selon le sondage, 9 255 personnes n'étaient pas dans ce cas à l'exhaustif. Comme la confusion de la

localisation se fait avec Mulhouse, il n'est pas étonnant que l'on trouve dans la colonne Tx1 le contrecoup de cette anomalie : 2 388 personnes se rendraient chaque jour à Mulhouse selon l'exhaustif, mais non selon le quart.

4) Résider à Strasbourg

Le groupe des actifs résidant à Strasbourg correspond à la sommation de deux des sous-groupes étudiés précédemment (les R & T plus les sortants). On obtient donc un tableau pour la population active résidente voisin des précédents.

Tableau 4 - Résider à Strasbourg
(sommation des tableaux 1 et 2)

Qualité des données	QT	origine	Cumul exhaustif	Cumul sondage	
				redressé	brut
cohérence	99	3/4	52 358	/	/
		3/4	54 363	/	/
		1/4	21 654	24 763	6.766
		1/4	19 819	71 653	19 819
incohérence	00	1 / 4	181	661	181
	04	1 / 4	334	1 204	334
	10	1 / 4	818	2 885	818
Totaux			**101 168**	**101 166**	

On remarque que l'écart entre les deux totaux est négligeable. La sous-estimation au sondage des R & T est exactement compensée par la surestimation des sortants.
On pourrait calculer également des taux d'incohérence, à l'instar de ce qui a été fait plus haut.

Tableau 4a- des résultats pour quelques communes

communes		incohérences		effectifs de référence			Proportion d'erreurs %	
		e1	e4	rf1	rf4	dif.	tx1	tx4
67482	Strasbourg	1 133	4 750	27 918	101 166	- 2	4,8	4,7
68224	Mulhouse	1 012	3 577	11 463	41 205	91	**8,8**	**8,7**
68300	Sausheim	41	164	601	2 404	63	6,8	6,8
74011	Annecy le V.	596	2 384	2 046	8 184	- 49	**29,1**	**29,1**
75056	Paris	10 312	39 691	253 803	975 608	- 491	4,1	4,1

Noter que la différence (dif) correspond à l'écart entre les valeurs du quart (redressées) et les valeurs de l'exhaustif. Les différences ne sont jamais importances du fait de la nature du tirage des logements (tirage équilibré).

5 - Travailler à Strasbourg

De la même façon on compose un tableau par sommation des tableaux 1 et 3 (les R& T plus les entrants)

Tableau 5 - Travailler à Strasbourg

Qualité des données	QT	origine	Cumul exhaustif	Cumul sondage	
				redressé	brut
cohérence	99	3/4	52 358		
	19	3/4	54 363		
	00	1/4	21 654	71 286	21 654
	14	1/4	19 819	71 653	19 819
incohérence	00	1/4	336	1 370	407
	04	1/4	728	1 204	334
	10	1/ 4	818	2 091	645
Totaux			**150 076**	**147 604**	

Cette fois l'écart entre les résultats des deux traitements est très forte : le sondage donne 2 472 emplois de moins que l'exhaustif.

Tableau 5a - Quelques résultats par commune

communes		incohérences		effectifs de référence			Proportion d'erreurs %	
		e1	e4	rf1	rf4	dif.	tx1	tx4
67482	Strasbourg	1 882	4 665	43 355	147 604	2 2472	4,3	3,2
68224	Mulhouse	3 057	1 357	18 467	53 040	- 9 101	**16,6**	2,6
68300	Sausheim	57	9 367	1 251	13 318	9 148	4,6	**70,3**
74011	Annecy le V	605	567	2 643	8 068	- 2 236	**22,9**	7,0
75056	Paris	25 248	37 021	433 670	1 582 596	- 54 745	5,8	2,3
92026	Courbevoie	5 111	12 645	20 798	72 672	- 6 931	**24,6**	**17,4**

Les anomalies apparaissent donc très clairement.

6) Regroupement des résultats

Les deux tableaux qui suivent rassemblent les différentes composantes des cinq tableaux précédents. On distingue cette fois un tableau pour les observations de l'exploitation principale (tableau 6a) et un autre pour l'exploitation complémentaire. Ces données vont être à la base des indicateurs permettant d'alerter sur les anomalies.

Tableau 6 - Les observations à l'exhaustif

QQ		QT	actifs	emplois	Résident et travaillent	Sortants	Entrants
0		09	18 887	52 358		18 887	52 358
(les 3/4 seulement)		19	54 363	54 363	54 363		
1		00	6 766	21 654		6 766	21 654
(cohérence)		14	19 819	19 819	19 819		
9		00	181	336		181	336
(incohérences)		04	334	728		334	728
		10	818	818	818		
regroupements							
cohérence	**3/4**	QQ=0	73 250	106 721	54 363	18 887	52 358
	1/4	QQ=1	26 585	41 473	19 819	6 766	21 654
incohérences		QQ=9	1 333	1 882	818	515	1 064
TOTAL			**101 168**	**150 076**	**75 000**	**26 168**	**75 076**

Mode de lecture. Le total des actifs s'obtient en additionnant les données des 3/4 (R & T plus sortants), et les données du quart. En ce cas il faut non seulement distinguer entre les deux catégories, mais il faut tenir compte des incohérences.

Tableau 7 - Les observations au sondage (après redressement)

QQ	QT	actifs	emplois	Résident et travaillent	Sortants	Entrants
0 (les 3/4 seulement)	09	.	.	.	.	
	19	.	.	.	.	.
1 (cohérence)	00	24 763	71 286		24 763	71 286
	14	71 653	71 653	71 653		
9 (incohérences)	00	661	1 370		661	1 370
	04	1 204	1 204	1 204		
	10	2 885	2 091		2 885	2 091
regroupements						
Cohérence	QQ=1	96 416	142 939	71 653	24 763	71 286
Incohérences	QQ=9	4 750	4 665	1 204	3 546	3 461
TOTAL		96 416	142 939	71 653	24 763	71 286

Dans ce tableau, les 3/4 est vide par construction.

7) Les indicateurs

Une fois les données élémentaires rassemblées, il faut déterminer des indicateurs pertinents. On a déjà, à chaque étape, calculé des taux d'erreurs.

On propose également de déterminer des indicateurs normés. En effet, qu'ils s'expriment en valeurs absolues ou en valeurs relatives, les écarts entre communes ne sont pas faciles à interpréter en raison de l'importance de la population qui diffère très fortement d'un cas à l'autre. On cherche donc à s'affranchir de la taille des communes.

Pour ce faire, on admet que les distributions des écarts entre les deux sources suivent une loi normale dont l'écart type est donné par $\sigma = 2\sqrt{x}$. Par conséquent, 95 % de chances que la valeur inconnue se situe dans un intervalle de largeur égale à $\pm 1{,}96.\sigma$, soit approximativement $\pm 4\sqrt{x}$ (Cette formule est donnée dans le fascicule vert de diffusion des résultats du RP1999. Dans la formule, x est l'effectif observé (ici, par convention on choisit de prendre la moyenne des effectifs : exhaustif + sondage).

Tableau 8 - Les indicateurs pour la commune de Strasbourg

	Actifs	Emplois	R & T	sorties	entrées
Effectifs totaux	101 168	150 076	*75 000*	*26 168*	*75 076*
Écarts (S-E)	- 2	- 2 472	*- 2 413*	*2 141*	*- 329*
Écart relatif (%)	0,0	- 1,7	*- 2,9*	*7,9*	*- 0,4*
Ecart normé	0,00	- 3,20	*- 3,94*	*6,49*	*- 0,60*
Incohérences (e1.)	4 750	4 665	*1 204*	*3 546*	*3 461*
Incohérences (e4)	1 333	1 882	*818*	*515*	*1 064*
Taux d'erreurs tx1 %	4,8	4,3	*4,0*	*7,1*	*4,7*
Taux d'erreur tx4 %	4,7	3,2	*1,7*	*12,5*	*4,6*

On rappelle que l'égalité quasi parfaite entre exhaustif et sondage pour la population active cache des écarts très importants dès lors que l'on effectue une partition de l'ensemble, en distinguant ceux qui sortent et ceux qui restent dans la commune pour y travailler.

Remarquer que l'écart relatif serait probablement plus grand encore si l'on décomposait davantage cette population active : en distinguant les jeunes des autres plus âgés, les femmes des hommes, etc.

8 - quelques résultats extrêmes

On présente quelques résultats pour des communes ayant un écart normé sur l'emploi au moins égal supérieur à 10 écarts types. Il s'agit donc forcément de cas statistiquement anormaux.

E1 et E4 sont les effectifs correspondant aux incohérences, RF1 correspondent aux effectifs de référence, TX1 et TX4 correspondent aux proportions d'anomalies (voir les tableaux 1, 2 ou 3).

dc	nom	tt	tx1	tx4	a1	a4	rf1	rf4
06018	BIOT	7,0	20,8	37,6	152	1371	730	3650
08480	VILLERS-SEMEUSE	17,2	10,2	56,7	52	2090	508	3689
20342		-8,9	55,8	3,2	416	11	745	340
22070	GUINGAMP	-7,3	20,5	4,9	456	235	2223	4825
29105	LANDIVISIAU	-6,5	20,2	2,7	357	143	1766	5389
30028	BAGNOLS-SUR-CEZ	-9,1	19,2	2,9	487	237	2540	8230
30081	CHUSCLAN	10,6	10,5	37,6	80	1624	765	4322
45055	BRICY	-19,9	71,1	0,7	409	4	575	542
45285	SAINT-JEAN-DE-L	-6,7	19,2	6,8	443	504	2309	7431
50041	BEAUMONT-HAGUE	-22,6	49,6	3,8	801	122	1615	3246
50242	HERQUEVILLE	38,4	4,9	88,4	5	2921	103	3304
59165	CUINCY	18,1	4,3	40,7	61	3227	1409	7929
59178	DOUAI	-9,7	15,7	4,1	1270	1007	8071	24800
59256	FRETIN	28,8	8,1	77,7	19	2408	235	3098
59343	LESQUIN	-13,4	27,7	3,5	851	291	3077	8277
59603	TRITH-SAINT-LEG	13,0	4,2	41,9	26	1652	622	3946
59606	VALENCIENNES	-7,9	11,7	3,3	991	931	8461	28594
68224	MULHOUSE	-19,0	16,6	2,6	3057	1357	18467	53040
68300	SAUSHEIM	50,8	4,6	70,3	57	9367	1251	13318
69123	LYON	-6,3	6,7	4,8	4518	11831	67640	244108
74011	ANNECY-LE-VIEUX	-11,7	22,9	7,0	605	567	2643	8068
75056		-21,6	5,8	2,3	25218	37021	433670	1582596
76305	GONFREVILLE-L'O	6,8	8,4	22,2	160	1898	1899	8554
76341	HARFLEUR	-9,8	35,7	10,3	389	309	1089	3013
77111	CHESSY	21,0	4,5	40,5	84	4052	1853	9997
77288	MELUN	-7,3	12,6	6,4	896	1503	7135	23564
77296	MOISSY-CRAMAYEL	-12,2	35,1	8,2	575	363	1638	4420
77384	REAU	26,2	6,6	67,6	26	2910	391	4307
78029	AUBERGENVILLE	14,6	4,8	28,3	94	2754	1955	9745
78073	BOIS-D'ARCY	-8,3	30,1	7,4	344	246	1144	3320
78238	FLINS-SUR-SEINE	-23,1	58,8	4,7	654	84	1112	1769
78297	GUYANCOURT	9,5	8,1	20,4	439	4844	5417	23781
78423	MONTIGNY-LE-BRE	10,8	13,9	25,1	739	5868	5317	23361

78621	TRAPPES	-14,9	24,0	7,0	1418	1254	5899	18026
78644	VERRIERE	-9,5	31,1	8,9	338	243	1086	2716
91182	COURCOURONNES	27,3	9,8	58,0	82	4052	838	6983
91228	EVRY	-19,2	22,5	5,7	2270	1832	10102	31873
91272	GIF-SUR-YVETTE	-10,5	28,7	8,7	613	562	2135	6455
91340	LISSES	10,0	7,3	27,9	99	1859	1363	6666
91479	PARAY-VIEILLE-P	7,8	6,5	23,6	83	1384	1279	5859
91521	RIS-ORANGIS	6,2	6,8	19,9	112	1468	1647	7371
91534	SACLAY	11,3	5,1	32,6	60	2070	1183	6358
91661	VILLEBON-SUR-YV	25,4	5,9	51,7	56	3723	954	7198
91692	ULIS	-12,6	22,4	5,0	1237	869	5517	17254
92026	COURBEVOIE	-12,6	24,6	17,4	5111	12645	20798	72672
93005	AULNAY-SOUS-BOI	-8,6	16,8	6,9	1414	1949	8394	28357
93070	SAINT-OUEN	9,2	6,0	14,6	423	4330	7058	29715
93073	TREMBLAY-EN-FRA	24,8	6,5	40,9	154	5921	2386	14471
93078	VILLEPINTE	7,3	7,5	20,8	207	2528	2742	12167
94021	CHEVILLY-LARUE	25,4	6,0	48,5	97	4708	1620	9699
94065	RUNGIS	-15,4	23,9	6,1	1603	1236	6695	20179
95527	ROISSY-EN-FRANC	-15,6	15,1	4,5	2483	2495	16490	55148

9 - L'activité économique

Tous ces résultats, pour autant qu'ils correspondent au bloc du 1/4, peuvent être déclinés selon les différentes variables du fichier (sexe, âge, diplôme, etc). On présente quelques résultats par activité économique.

Le tableau donne, dans la NAF60, les anomalies selon la fréquence décroissante.

Tableau 10

Activité économique	Code	Taux d'erreur	effectif de référence	erreurs
Cokéf. raffin. nucléaire	23	**25,0**	22 861	5 711
Hycrocarbures (annex.)	11	**22,7**	5 581	1 267
Exploitation forestière	02	**20,5**	43 240	8 864
construction	45	15,1	1 331 542	200 491
Transports par eau	61	14,3	14 723	2 109
Automobile	34	11,5	273 554	31 490
informatique	72	11,3	244 121	27 555

Chimie	24	10,4	270 141	27 990
Métallurgie	27	10,2	127 665	13 033
Act. récréatives	92	10,1	376 315	38 164
Comm. de gros	51	10,1	976 146	98 165
div. matériel de transport	35	9,7	144 307	14 066
Div ind. extractives	14	9,7	28 673	2 780
Services aux entreprises	74	9,4	1 835 954	173 119
Transports aériens	62	9,4	65 442	6 137
Fabr. mach. bureau et inform.	30	8,9	34 767	3 092
Location	71	8,8	63 858	5 614
Transports terrestres	60	8,7	662 927	57 473
……..				
Moyenne		6,5	22 862 088	1 474 818

Les deux tableaux suivants donnent des résultats plus détaillés. Ils sont classés par NAF et par département. Il permettent d'identifier presque sûrement les noms des communes faisant l'objet d'une anomalie de chiffrement. Toutefois il n'est pas possible de déterminer si l'erreur est due à l'exhaustif ou au sondage. La colonne a1' correspond aux erreur a1 multipliées par 4 (inverse du taux de sondage) de façon à rendre comparable les efectifs anormaux de l'exhasutif et du sondage (rappelons une fois encore que le coeffciient n'est pas toujours égal à cause des communes ou parties de commune traitées exhas Ustivement dans l'exploitatin complémentaire.

Le premier tableau (10A) fournit la liste des cas où la somme des erreurs est supérieure à 1000. Le plus souvent, pour un même département on a un total des effectifs de la colonne a1' à peu près égal au même total pour a4. Ainsi les couple Beaumont-Hague / Herqueville, le Tréport/Mers-les Bains, etc. Les communes jumelles sont le plus souvent limitrophes.

Les second tableau (10B) est présenté de la même façon, mais il ne concerne que les activités de construction de matériel de transport. Le seuil est abaissé à 200.

Tableau 10 A : les principales anomalies par commune

NAF	dc	nom	a1’	a4	rf1	rf4
233Z	50242	HERQUEVILLE	0	2841	78	3144
233Z	50041	BEAUMONT-HAGUE	2496	0	679	217
261E	76711	TREPORT	0	1025	3	1037
261E	80533	MERS-LES-BAINS	1060	0	286	78
271Z	59183	DUNKERQUE	1592	12	459	223
271Z	59271	GRANDE-SYNTHE	0	1474	1257	6010
275A	08480	VILLERS-SEMEUSE	0	1650	115	2044
275A	08040	AYVELLES	1096	0	286	42
322B	78423	MONTIGNY-LE-BRE	44	1020	69	1243
322B	78073	BOIS-D'ARCY	1044	0	317	218
341Z	59178	DOUAI	3068	4	815	169
341Z	59165	CUINCY	0	2865	834	5760
341Z	68224	MULHOUSE	8696	8	2352	543
341Z	68300	SAUSHEIM	0	8323	381	9544
341Z	78029	AUBERGENVILLE	4	2311	978	5919
341Z	78238	FLINS-SUR-SEINE	2164	0	609	254
341Z	93005	AULNAY-SOUS-BOI	2380	4	1005	1575
341Z	93070	SAINT-OUEN	8	2274	142	2819
343Z	25580	VALENTIGNEY	0	1081	282	2143
353A	77384	REAU	0	2435	243	3371
353A	77296	MOISSY-CRAMAYEL	1868	0	540	277
353B	31555	TOULOUSE	88	1620	2154	9866
353B	31149	COLOMIERS	1064	20	486	891
513A	94065	RUNGIS	1456	49	688	1249
513A	94021	CHEVILLY-LARUE	0	1399	13	1445
621Z	75056	Paris	752	272	1355	4667
621Z	95527	ROISSY-EN-FRANC	448	1370	6020	23352
634C	93073	TREMBLAY-EN-FRA	0	1269	60	1503
634C	95527	ROISSY-EN-FRANC	1364	46	965	2407
642B	92062	PUTEAUX	284	1296	581	3258
642B	92026	COURBEVOIE	1208	132	666	1519
651C	75056	Paris	628	931	15330	58080
651C	92062	PUTEAUX	704	2716	1643	8308
651C	92026	COURBEVOIE	2948	469	1496	3361
660A	92026	COURBEVOIE	460	639	465	1943
660A	92062	PUTEAUX	584	443	339	1182
721Z	75056	Paris	2972	1242	4892	16929
721Z	92062	PUTEAUX	888	1025	1014	4040
721Z	92026	COURBEVOIE	1380	484	734	1995
722Z	75056	Paris	1632	624	3918	13974
722Z	92026	COURBEVOIE	504	675	609	2523
731Z	91534	SACLAY	4	1587	600	3857
731Z	91272	GIF-SUR-YVETTE	1424	0	543	709
741G	75056	Paris	1644	1024	7597	28379

741G	92026	COURBEVOIE	328	803	453	2212
741J	75056		1592	570	6815	24768
741J	78297	GUYANCOURT	20	1380	222	2224
741J	92026	COURBEVOIE	1280	1060	1782	6617
741J	92062	PUTEAUX	1032	1265	1178	4807
742C	75056	Paris	1172	595	3670	13350
742C	92026	COURBEVOIE	364	884	720	3307
742C	92062	PUTEAUX	908	240	468	1153
745B	69123	LYON	896	405	1279	4292
745B	75056	Paris	5028	1199	6473	20842
751A	75056	Paris	2112	2034	21713	81941
751A	92062	PUTEAUX	152	1129	709	3714
752C	20342	Paris	1536	1	476	93
752C	21355	LONGVIC	1252	0	348	125
752C	21231	DIJON	4	1240	352	2596
752C	27229	EVREUX	1172	19	400	408
752C	27234	FAUVILLE	0	1073	20	1138
752C	29105	LANDIVISIAU	1004	5	347	371
752C	33281	MERIGNAC	1328	32	580	904
752C	33063	BORDEAUX	20	1268	675	3780
752C	45055	BRICY	1564	4	519	411
752C	45234	ORLEANS	0	1516	355	2825
752C	57463	METZ	88	934	991	4621
752C	68205	MEYENHEIM	1044	0	339	270
851A	38185	GRENOBLE	992	45	766	2018
851A	69123	LYON	1188	847	4980	18199
851A	69029	BRON	108	967	801	3916
851A	75056	Paris	1152	991	16837	63713
851A	93005	AULNAY-SOUS-BOI	1056	16	460	752
851A	93078	VILLEPINTE	16	990	112	1398

Tableau 10B - Fabrication de matériel de transport

	NAF	dc	nom	a1'	a4	rf4
Véhicules automobiles						
	341Z	25547	SOCHAUX	68	712	15290
	341Z	25388	MONTBELIARD	632	56	908
	341Z	35066	CHARTRES-DE-BRE	4	558	8569
	341Z	35238	RENNES	532	23	2112
	341Z	57677	TREMERY	0	609	2905
	341Z	57283	HAGONDANGE	356	0	17
	341Z	57193	ENNERY	208	8	53
	341Z	59178	DOUAI	3068	4	169
	341Z	59165	CUINCY	0	2865	5760
	341Z	59348	LIEU-SAINT-AMAN	0	848	3567
	341Z	59092	BOUCHAIN	412	0	28
	341Z	59313	HORDAIN	328	0	26
	341Z	59392	MAUBEUGE	0	210	2817
	341Z	62276	DOUVRIN	0	284	4614
	341Z	68224	MULHOUSE	8696	8	543
	341Z	68300	SAUSHEIM	0	8323	9544
	341Z	68154	ILLZACH	736	0	56
	341Z	70550	VESOUL	4	216	2139
	341Z	75056	Paris	184	52	2219
	341Z	76660	SANDOUVILLE	4	756	5762
	341Z	76351	HAVRE	672	0	203
	341Z	76178	CLEON	0	237	4653
	341Z	78029	AUBERGENVILLE	4	2311	5919
	341Z	78238	FLINS-SUR-SEINE	2164	0	254
	341Z	78498	POISSY	36	214	5285
	341Z	78297	GUYANCOURT	12	213	5079
	341Z	93005	AULNAY-SOUS-BOI	2380	4	1575
	341Z	93070	SAINT-OUEN	8	2274	2819
Equipements pour l'automobile						
	343Z	12202	RODEZ	360	0	49
	343Z	12176	ONET-LE-CHATEAU	0	336	1430
	343Z	25580	VALENTIGNEY	0	1081	2143
	343Z	25367	MANDEURE	872	0	63
	343Z	25031	AUDINCOURT	284	49	1638
	343Z	49007	ANGERS	120	113	540
	343Z	59603	TRITH-SAINT-LEG	0	948	1241
	343Z	59606	VALENCIENNES	800	0	91
Bâtiments de guerre						
	351A	29019	BREST	8	198	4519
	351A	44074	INDRE	236	0	64
	351A	44101	MONTAGNE	0	231	1227
	351A	56098	LANESTER	260	0	58
	351A	56121	LORIENT	4	204	2344

Moteurs pour aéronefs						
	353A	76305	GONFREVILLE-L'O	0	409	995
	353A	76341	HARFLEUR	336	0	8
	353A	77384	REAU	0	2435	3371
	353A	77296	MOISSY-CRAMAYEL	1868	0	277
	353A	77288	MELUN	624	0	161
	353A	91174	CORBEIL-ESSONNE	4	645	1859
	353A	91228	EVRY	572	0	134
Cellules d'aéronefs						
	353B	13054	MARIGNANE	4	228	4998
	353B	31555	TOULOUSE	88	1620	9866
	353B	31149	COLOMIERS	1064	20	891
	353B	31069	BLAGNAC	416	115	2612
	353B	33063	BORDEAUX	40	426	1049
	353B	33167	FLOIRAC	420	0	65
	353B	35228	PLEURTUIT	0	228	415
	353B	35093	DINARD	220	0	20
	353B	44184	SAINT-NAZAIRE	20	202	2381
	353B	65284	LOUEY	0	418	756
	353B	65344	OSSUN	240	0	28
	353B	83137	TOULON	4	326	380
	353B	83049	CUERS	292	0	403
	353B	93013	BOURGET	192	24	391
Lanceurs et engins spatiaux						
	353C	33449	SAINT-MEDARD-EN	192	28	1374
	353C	77384	REAU	0	385	401
	353C	77288	MELUN	376	0	22
Motocycles						
	354A	02659	ROUVROY	4	690	1316
	354A	02691	SAINT-QUENTIN	632	4	44

II - D'autres résultats

1 - La résidence antérieure.

Alors que les données sont cohérentes par construction pour toute la population née avant le 1er janvier 1999, il apparaît une différence systématique de codification entre les deux exploitations pour les enfants nés au cours de la dernière période intercensitaire.
Cette différence est de nature conceptuelle. En effet, pour les enfants nés au cours de la période intercensitaire, il ne saurait y avoir de résidence antérieure (à la date du précédent recensement). Pour ne pas laisser à blanc l'information, on est donc amené à décider d'une convention. Celle-ci n'a pas été la même dans les deux phases de traitement du recensement.

- dans l'exploitation principale la résidence antérieure affectée à l'enfant est celle de la personne de référence du *ménage* (c'est la seule façon de faire puisque le concept de famille n'est pas pris en compte dans l'exploitation).
- dans l'exploitation complémentaire la résidence antérieure affectée à l'enfant est celle de la mère (éventuellement le père en cas de monoparentalité). Cette façon de codifier n'est possible que lorsque a été affecté un code lien familial à chaque individu. La phase de composition des familles n'a pas été prévue dans l'exploitation principale.

Aux recensements antérieurs (jusqu'en 1990 inclus), la question de la divergence de codification de la résidence antérieure de l'enfant ne se posait pas puisque la codification correspondante ne se faisait que dans le cadre de l'exploitation complémentaire (qui comportait un code lien familial).

Cette façon différente de codifier les enfants entraîne des divergences importantes dans les résultats, comme le montre le tableau qui suit, donnant le taux d'anomalies par âge de l'enfant. En moyenne le taux d'incohérence est de 26,9 % pour les enfants nés au cours de la période 1990-1999. Le lien entre le taux d'anomalie et l'âge de l'enfant est net car la propension à migrer des parents est importante dans la période précédant et suivant la naissance des enfants.

Depuis 2004, la localisation des enfants de moins de 5 ans n'est pas codifiée, ce qui règle la question pour l'exploitation des recensements de la nouvelle génération, mais qui ne la règle que partiellement quand il s'agit du suivi historique des mouvements migratoires. Il est vrai qu'il demeure un problème qui rend les comparaisons temporelles délicates, à savoir la durée inégale des périodes intercensitaires.

Tableau 11 Taux d'incohérences sur la résidence antérieure, par âge

Age de l'enfant	Taux d'incohérences %
000	51,2
001	48,7
002	43,5
003	37,2
004	31,4
005	25,6
006	20,3
007	14,9
008	10,4
009	7,7
moyenne	**26,9**

Le taux d'incohérence est calculé en divisant, par âge, et pour l'ensemble de la métropole, le nombre de cas où la commune de résidence antérieure de l'enfant est différente dans les deux traitements.

2 - Code nature de l'emploi (code EMPL)

Les personnes actives ayant un emploi sont classées selon la nature de leur emploi dans une nomenclature en dix modalités (dont trois pour les non salariés).
Globalement entre les données des deux exploitations les écarts sont les suivants (base exhaustif)

Tableau 12a : divergences selon les 10 modalités du code EMPL

Catégorie d'emploi	code	effectifs			*Présents seulement*	
		Exhaustif	Sondage	Différence	à l'exhaustif	Au quart
Apprentis	11	83 076	82 956	- 120	4 009	3 889
Intérimaires	12	109 765	118 387	8 622	6 823	15 445
Empl. aidés	13	137 969	138 970	1001	9 157	10 158
Stagiaires rém.	14	44 527	44 302	- 225	3335	3 110
CDD	15	485 952	486 628	676	37 126	37 802
Fonction publique	16	1 119 688	1 115 030	- 4 658	70 619	65 961
CDI	17	3 629 715	3 671 215	41 500	109 106	150 606
Indépendants	21	382 090	372 599	- 9 491	44 508	35 017
Employeurs	22	339 268	317 073	- 22 195	34 186	11 991
Aides familiaux	23	74 490	59 380	- 15 110	15 202	92
Ensemble		**6 406 540**	**6 406 540**	**0**	**334 071**	**334 071**

Mode de lecture : Pour les salariés de la fonction publique (code EMPL=16), l'exploitation principale donne 1 119 688 d'enregistrements, l'exploitation par sondage en donne 4 658 de moins. L'écart paraît faible. Mais, lorsqu'on regarde plus en détail, on constate que 70 619 individus, soit, 6,3%, ont été codifiés comme fonctionnaires à l'exhaustif mais non dans le sondage, cependant que 65 961 ont été codifiés de façon inverse. Rien ne dit que ces erreurs de codification se répartissent uniformément en cas de déclinaison des résultats.

Tableau 12b - indicateurs

Catégorie d'emploi	Code EMPL	anomalies apparentes	Actifs présents seulement	
			À l'exhaustif	Au quart
		%	%	%
Apprentis	11	- 0,1	4,8	4,7
Intérimaires	12	7,95	6,2	14,1
Empl. aidés	13	0,7	6,6	7,4
Stagiaires rém.	14	- 0,5	7,5	7,0
CDD	15	0,1	7,6	7,8
Fonction publique	16	- 0,4	6,3	5,9
CDI	17	1,1	3,0	4,2
Indépendants	21	- 2,5	11,6	9,2
Employeurs	22	- 6,5	10,1	3,5
Aides familiaux	23	- 20,3	20,4	0,1
Ensemble	.	0,00	5,2	5,2

Les données de base sont celles du tableau précédent. Les anomalies apparentes sont obtenues en divisant l'écart par la moyenne des deux traitements. Les deux autre pourcentages sont obtenu sen divisant les personnes présentes à l'exhaustif (resp. au quart)par l'effectif total correspondant.

Si l'on considère l'ensemble de
s actifs, les divergences sur la catégorie d'emploi touchent 5,2 % des cas. Elle sont particulièrement importantes pour les aides familiaux puisque 15 202 personnes ont été codifiées comme telles à l'exhaustif (et non au quart) contre seulement 92 au quart (et non à l'exhaustif). Par rapport au total des aides familiaux, cela fait 20% des cas qui seraient codifiés ainsi à tort à l'exhaustif.

Le tableau suivant présente les seuls actifs ayant été codifiés comme fonctionnaires à l'une ou l'autre des deux exploitations. Les divergences sont importantes (des CDI notamment), mais aussi des dissymétries très fortes, notamment les personnes déclarées - probablement à tort - « fonctionnaires » à l'exhaustif et « aides familiales » au quart.

Tableau 12 c - effectifs et écarts selon les modalités du code EMPL

Cas particulier des **fonctionnaires (EMPL=16**)

catégories	modalités		Anomalies apparentes	Présents seulement	
	Réf.	Ech*.		À l'exhaustif	Au quart
Apprentis	16	11	- 4	213	209
Intérimaires	16	12	- 291	1 137	846
Empl. aidés	16	13	- 252	1 585	1 333
Stagiaires rém.	16	14	- 45	380	335
CDD	16	15	- 1 235	6 247	5 012
Fonction publique	16	16	0	/	/
CDI	16	17	- 8 791	55 326	46 535
Indépendants	16	21	572	5 115	5 687
Employeurs	16	22	3 772	611	4 383
Aides familiaux	16	23	1 616	5	1 621
Total	16	.	- 4 658	70 619	65 961

** modalité avec laquelle il y a eu substitution*

Mode de lecture : la ligne « total » correspond à la surestimation de l'exhaustif pour les « fonctionnaires » (cf ligne '16' du tableau précédent). Cet écart se décline lui-même selon les 9 autres modalités de la nomenclature. Ainsi la ligne « aides familiaux » contribue à raison de 1616 personnes. Ce chiffre est la somme algébrique de deux éléments : 5 personnes classées comme fonctionnaires à l'exhaustif ont été intégrées dans la modalité « aides familiaux » au quart, mais 1621 se sont trouvés dans la situation inverse.

3 - Les frontaliers[4]

L'importance numérique, et surtout les variations parfois brutales d'effectifs qui caractérisent cette catégorie de la population, jouent un rôle essentiel dans l'équilibre du marché du travail de certaines régions, et plus particulièrement de certaines zones d'emploi. On aurait aimé être en mesure de fournir des données de qualité tant structurelles que conjoncturelles pour les régions proches de la frontière. On ne doit pas oublier que, en l'absence de sources administratives homogènes, les recensements de la population forment la source principale d'information sur les travailleurs frontaliers.

En 1999, combien de travailleurs frontaliers ?

Si on se réfère aux chiffres de l'exploitation principale, on dénombre 278 922 personnes ayant déclaré une activité en dehors de l'hexagone (sans les DOM-TOM). Les chiffres de l'exploitation complémentaire sont plus faibles, 249 313, soit près de 30 000 de moins. La divergence (environ 10%), n'est pas négligeable.

Tableau 13 - le travail hors métropole : données d'ensemble (RP1999):

Qulaité des données	origine	Cumul exhaustif	Cumul sondage	
			redressé	brut
cohérence	3/4	**201 053**		
	1/4	**70 067**	247 313	70 067
	Total	**271 120**	247 313	
incohérence	1/4	**8 802**	31 326	8 802
	1/4	499	**1 178**	499
Totaux		**279 922**	**248 491**	79 308

Mode de lecture (comparable à celui des tableaux des pages précédentes)

***En ligne**. On distingue deux cas de figure, selon qu'il y a ou non cohérence dans la déclaration du caractère frontalier.*
*Quand il y a **cohérence,** deux cas ont envisagés. Le premier concerne les situations où la personne figure uniquement dans l'exhaustif (origine=1/1), le second quand l'individu appartient au champ du sondage (1/4).*
*Quand il y a **incohérence**, il s'agit des individus du quart pour lesquels la déclaration diverge selon les traitements.*

***En colonne**. Les cumuls se font soit sur la variable SC1 (cumul exhaustif), soit sur la variable SC4 (cumul sondage).*
Le total pour l'exhaustif comprend trois catégories : les personnes des 3/4 : 201 053 frontaliers), celles du 1/4 pour lesquelles il y a cohérence (70 067), et les personnes du 1/4 pour lesquelles il y a incohérence (8 802 personnes). Au total cela fait 279 922 personnes.
Le total pour le sondage peut être vu de deux façons.

[4] *Ou plus précisément les personnes ayant déclaré un lieu de travail en dehors de la France métropolitaine.*

D'une part en nombre d'individus dans le fichier (colonne « brut »), à savoir les 70 067 individus cohérents, et les 499 personnes qui ne le sont pas (car non frontalières au sens de l'exhaustif). Le total n'a pas d'intérêt.
D'autre part, l'estimation du nombre de frontaliers au sondage. Ce sont les individus pondérés (multipliés par 4 à peu près). De sorte que le total des frontaliers au 1/4 est la somme de deux éléments : les cohérents (247 313) et les incohérents (1 178).

On constate que pour beaucoup le lieu de travail a souvent été codifié de façon incohérente. Sur un total de 79 308 enregistrements dans le fichier du quart, concernés par le travail hors de la frontière, on a dénombré

- 8 802 individus frontaliers à l'exhaustif seulement, soit 12,6%
- 499 codés non frontaliers à l'exhaustif et frontaliers au quart, soit seulement soit 0,7%.

Il apparaît donc une grande dissymétrie dans les anomalies de codification.

Les conséquences de telles divergences sont importantes pour les territoires qui sont confrontés au problème de l'emploi au-delà de la frontière.

Tableau 14 - les divergences par département

Départements	Effectifs cohérents	Présents seulement à l'exhaustif e1	Présents seulement au quart e4
Moselle	48126	3320	252
Haut-Rhin	38945	1733	270
Haute-Savoie	35715	1576	237
Bas-Rhin	29084	2159	277
Alpes-Maritimes	19472	1350	173
Ain	15313	709	204
Meurthe-et-Moselle	13433	1005	72
Nord	12875	1922	64
Doubs	9334	557	72
…			
Paris	2283	1522	8
Hauts-de-Seine	716	670	4
Gironde	403	395	0

Un regard historique 1968- 1999

Il faut bien reconnaître que l'imprécision des données sur les frontaliers dans les recensements n'est pas nouvelle. Le tableau qui suit donne un certain nombre de séries temporelles issues des exploitations par sondage pour la période allant de 1968 à 1999.

Tableau 15 - le travail hors métropole de 1968 à 1999

recensement		RP68	RP75	RP82	RP90	RP99
Taux de sondage		1/4	1/5	1/4	1/4	1/4
Ensemble (1)	1	42 372	95 175	104 648	199 888	249 835
Frontaliers (2)	***2***	***41 464***	***90 960***	***103 232***	***175 944***	***225 121***
Autres cas (3)	3	908	4 215	1 416	23 944	24 714
Dont ceux travaillant dans les cinq pays limitrophes Total (4)	4	776	3 485	1 116	17 704	6 279
dont Île-de-France (5)		0	835	68	4292	1513
dont Bretagne (6)		0	110	0	1512	185
dont Aquitaine (7)		0	80	0	1016	162
Dont autres pays (8)	5	132	730	300	6 240	18 435

Afin de montrer la difficulté de comparer dans le temps des données lorsque les modes de codification divergent, différentes partitions des effectifs ont été opérées comme suit:

Ligne 1 : ensemble des personnes actives occupées ayant été codifiées sans indication d'une commune de travail en métropole.

Ligne 2 : **frontaliers, définis en associant** des pays de travail à des départements limitrophes, ce qui permet de disposer de séries plus homogènes, donc plus proches de la notion de frontaliers. Les départements associés sont, pour la Belgique 02 - 08 - 25 - 54 - 55 – 59, pour le Luxembourg 08 - 54 - 55 – 57, pour l'Allemagne 54 - 57 - 67 -68, pour la Suisse 01 - 25 - 39 - 67 - -68 - 74 - 90 et pour Monaco 06.
Ligne 3 : personnes travaillant dans les cinq pays désignés ci-dessus et résidant hors des départements listés (ne sont donc pas considérés comme frontaliers).
Ligne 4 : Personnes distantes déclarées comme travaillant dans les cinq pays limitrophes.
Ligne 5 à 7 : quelques départements de résidence éloignés.
Ligne 8 : personnes travaillant dans un autre pays que les cinq retenus.

On constate des ruptures de série très fortes d'un recensement à l'autre. Par exemple, le nombre de « frontaliers » lointains travaillant dans les cinq pays limitrophes a connu des fluctuations importantes : 3485 en 1975, puis 1116 en 1982, puis 17704 en 1990, puis 6279 en 1999.

L'accroissement du nombre de personnes travaillant dans un autre pays que les cinq retenus de 1990 à 1999 est lié l'émergence de nouveaux pays frontaliers, malgré la distance séparant le lieu de domicile du lieu de travail : l'Angleterre (beaucoup de résidents de départements du sud indiquent ce pays comme lieu de travail), les Pays-Bas, l'Italie et l'Espagne, etc. Par contre, les travailleurs frontaliers vers l'Espagne et l'Italie ont toujours été peu nombreux.

4) Des erreurs communes aux deux traitements.

Il existe des erreurs de collecte qui auraient sans doute pu être aisément corrigées si un traitement statistique ad hoc avait été effectué. On voudrait évoquer quelques cas assez graves pour perturber des indicateurs statistiques dans le cadre d'une analyse territoriale.

a) Le cas de la commune Millizac.

Millizac est une petite commune du Finistère de près de 3000 habitants. Si l'on devait en croire les fichiers du recensement de 1999, cette commune aurait « exporté » plus de mille de ses habitants au bénéfice de la commune d'Ivry-sur- Seine au cours de la seule période 1990-1999. On constate que ces « migrants » seraient en fait des personnes âgées, qui résideraient dans une maison de retraite de la banlieue parisienne. Il s'agit donc de personnes hors résidence principales. Est-ce une anomalie due à un défaut de collecte, ou à un défaut de traitement ? Toujours est-il que de telles erreurs sont fâcheuses et perturbent les indicateurs statistiques.

En effet, la structure par âge des « migrants » comprend en effet seulement personnes de moins de 50 ans, de et plus, de 80 ans et plus.

Tableau 16 : les départs de la commune d e Millizac, comparés à ceux du Finistère

Age	Départs de Millizac (autres que vers Ivry)	Départs vers Ivry	Autres départs du Finistère	en % des départs du Finistère
Moins de 25 ans	4	427	34 752	0,0
25 à 54 ans	4	553	48 433	0,0
55 à 64 ans	12	38	2 285	0,5
65 à 79 ans	260	24	2 673	8,8
80 à 89 ans	452	4	659	**40,5**
90 à 99 ans	464	4	246	**65,0**
100 ans et plus	20	0	8	**71,4**
Tous âges	1 216	1 050	89 056	1,3

Ces anomalies sont évidente dès lors que l'on travaille sur de petits territoires. Mais si l'on vient à travailler sur de plus grandes masses, au niveau du département par exemple, et si on s'intéresse aux migrations des personnes âges, il est clair que les indicateurs migratoires vont s'en trouver faussés.

b) des erreurs sur le lieu de naissance

En réalisant toute une série de pyramides d'âges sur les quartiers de Strasbourg, un collègue (Jacques Postic) m'avait fait part de son étonnement en voyant apparaître dans un « iris2000 » un énorme pic. Un examen détaillé tant de l'exploitation principale que de l'exploitation complémentaire a montré qu'il s'agissait de personnes inscrites dans un établissement de

formation professionnelle pour les jeunes adultes. Non seulement plus de 100 personnes avaient donc exactement la même date de naissance (année, mois, jour puisqu'elles étaient nées le), mais, en y regardant de plus près, on constatait que toutes avaient la même résidence antérieure (Wasselonne), une petite commune située à une vingtaine de kilomètres de Strasbourg. On retrouve là, en plus atténué, le même problème que pour Millizac.

D'une observation faite sur l'ensemble des « iris2000 » de la métropole, que ce soit à l'exhaustif qu'au quart, il apparaît que ce type d'anomalie n'est pas rare. On présente deux tableaux.

Voici une liste des principaux cas (commune et n° iris).

Tableau 17 - Liste des cas où l'effectif dépasse 100

Les iris2000 sont classés selon la catégorie de population, par ordre d'effectif décroissant. On donne les effectifs dans l'exhaustif (colonne exh) et dans le sondage (sdg).

dep	dcr	IRIS2000	CATP	ANAI	MNAI	JNAI	exh	sdg
Logement ordinaire								
13	13215	0603	01	976	11	17	317	81
13	13215	0603	01	955	11	17	167	43
Foyer de travailleur								
75	75113	5014	11	945	12	18	432	109
95	95219	0501	11	994	06	10	267	67
94	94041	0206	11	914	07	16	256	64
84	84092	0104	11	936	11	11	225	56
95	95018	0405	11	968	05	15	199	50
94	94068	0701	11	968	06	05	161	40
67	67482	2501	11	956	12	01	141	35
75	75113	5117	11	955	03	02	137	34
95	95063	0101	11	938	12	01	130	32
75	75119	7524	11	946	02	17	125	31
Cité universitaire								
14	14118	2001	12	974	11	06	195	48
91	91228	0110	12	976	05	29	193	48
92	92019	0111	12	977	10	04	180	44
14	14327	0401	12	980	08	15	177	44
68	68224	0801	12	980	05	20	147	37
06	06088	2102	12	976	04	06	131	33
13	13213	0902	12	978	02	18	129	32
Etablissement de soins								
41	41018	0601	14	908	05	07	163	41
Centre d'hébergement (courte période)								
13	13202	0403	16	991	05	24	296	74
Sans abri								
13	13202	0203	22	980	11	18	400	92
13	13202	0201	22	948	06	16	193	48

13	13202	0201	22	951	04	26	182	66
13	13205	0301	22	961	02	25	123	23
Centres de détention								
92	92050	0604	33	941	06	29	344	344
66	66136	1701	33	976	10	16	138	34

Catp : catégorie de population
Anai, Mnai, Jnai : année, mois, jourd e naissance
Exh : efectif à l'exhaustif
Sdg, effectif au sondage

Tableau 18 Même date de naissance et la même résidence antérieure*

dep	**dcr**	**IRIS2000**	**CATP**	**ANAI**	**MNAI**	**JNAI**	**dcra**	**sdg**
Logement ordinaire								
13	13215	0603	01	976	11	17	41018	81
13	13215	0603	01	955	11	17	41018	43
Foyer de travailleur								
94	94068	0701	11	968	06	05	94081	40
93	93066	0902	11	975	10	23	57287	39
67	67482	2501	11	956	12	01	68066	35
Cité universitaire								
14	14118	2001	12	974	11	06	61214	48
91	91228	0110	12	976	05	29	34172	48
14	14327	0401	12	980	08	15	14597	44
92	92019	0111	12	977	10	04	94046	44
75	75118	7027	12	977	03	16	9B221	40
06	06088	2102	12	976	04	06	06120	33
13	13213	0902	12	978	02	18	01397	32
Etablissement hospitalier								
41	41018	0601	14	908	05	07	36018	41
Centre d'accueil courte durée								
13	13202	0403	16	991	05	24	13215	74

Dcra : commune de résidence antérieure ;

Il arrive donc fréquemment que des migrations de longue distance portant sur des populations nombreuses soient complètement fictives.

Il est clair que des tests statistiques simples effectués très tôt dans la chaîne de traitement aurait permis de corriger à temps des erreurs qui persistent encore dans les deux fichiers.

III - Méthode

1 - Exploitation principale et complémentaire : un fichier fusionné

On rappelle que le recensement de 1999, le dernier de la série des recensements traditionnels a fait l'objet, comme les précédents, de deux exploitations informatiques successives à but statistique, sachant que la première exploitation, le dénombrement, est avant tout une exploitation à but légal. L'exploitation **principale** (dite aussi **exhaustive**) a porté sur *l'ensemble* des logements et des personnes y habitant. N'étaient cependant pas pris en compte les variables dont la codification est complexe. Ainsi, la catégorie socio-professionnelle et l'activité économique, mais aussi le lien de chaque individu avec la personne de référence de la famille quand il y en a une n'ont été exploitées que dans la seconde phase, dénommée exploitation **complémentaire** (appelée aussi, quoique abusivement exploitation **au quart**[5]).

Si la plupart des variables individuelles et toutes les variables liées au logement sont identiques dans les deux fichiers, ce n'est pas le cas pour un petit nombre de variables attachées à la personne : c'est le cas notamment de la catégorie d'emploi (code EMPL en dix positions) et du lieu de travail pour les actifs (la commune, voire le pays de travail). La seconde codification, au quart, est a priori de meilleure qualité puisqu'elle est conduite à l'aide de systèmes experts prenant en compte les caractéristiques de l'établissement employeur tel que connu par le fichier Sirène.

Les divergences, parfois importantes, qui apparaissent, contribuent donc à semer le doute chez les utilisateurs dès lors qu'ils sont confrontés à des résultats divergents, voire fortement contradictoires. En effet, les données issues des deux traitements continuent d'être diffusées simultanément (dans EDL), sans qu'il y ait, à ma connaissance, d'explication précise propre à aider les utilisateurs.

Un exemple : l'une des principales divergences repérées a trait à la localisation des salariés de l'usine dite de Peugeot-Mulhouse. En réalité, les quelque 13000 salariés en question travaillent à Sausheim, commune voisine de Mulhouse. Dans le cadre du traitement principal, la plupart de ces salariés ont été codifiés - à tort - avec pour lieu de travail la commune de Mulhouse (code 68224). En revanche, dans l'exploitation complémentaire, la localisation a été codifiée correctement (Sausheim n° 68300). Qui n'est pas informé de cette incohérence peut être amené à tirer des conclusions erronées d'importance, et ce d'autant plus que les localités de Mulhouse et de Sausheim appartiennent à certains regroupements communaux différents. On peut donc imaginer des batailles de chiffres entre élus défendant chacun leur position. À moins que - ce qui est le plus probable - ces élus se désintéressent finalement d'une information dénuée d'intérêt car trop éloignée de la réalité.
Et si l'on peut accepter l'idée que les agents territoriaux corrigeront d'eux-mêmes cette anomalie flagrante, il est moins sûr que les statisticiens, travaillant loin de l'Alsace, en feront

[5] En effet dans un certain nombre de communes, l'exploitation complémentaire a porté sur l'ensemble de la population.

d'eux-mêmes quand ils calculeront des indicateurs globaux d'évolution des navettes intercommunales.

Le piège est d'autant plus ennuyeux que les erreurs ne vont pas toujours dans le même sens. Ainsi les militaires de la base aérienne de Colmar-Meyenheim sont classés comme il faut dans le traitement principal, mais de façon erronée dans le traitement complémentaire. Or ces deux communes, quoique assez proches l'une de l'autre, appartiennent à deux zones d'emploi différentes (l'une dans celle de Guebwiller, l'autre dans celle de Colmar). Tel chargé d'études de la DR d'Alsace qui avait à étudier l'évolution du marché du travail de la zone d'emploi de Guebwiller a reconnu avoir perdu beaucoup de temps par manque d'information sur cette question (d'autant plus, que lorsque les chiffres des deux traitements se contredisent, la « *bonne réponse* » - celle que l'entourage apporte spontanément - est de dire que les meilleurs résultats sont ceux issus du sondage !). Affecter ou non au territoire de Guebwiller, qui compte environ 20 000 postes de travail, les quelque 600 militaires de la base aérienne, change le point de vue sur le dynamisme économique de la zone d'emploi : 3 points d'écarts sur une variation, ce n'est pas négligeable. C'est plus encore si l'on considère des sous-groupes (le sexe, l'âge, les catégories socioprofessionnelles ou les activités économiques). On peut arriver très vite à dire n'importe quoi ![6]

On est donc fondé à s'étonner que l'on n'ait pas cherché à donner aux utilisateurs les clés leur permettant d'éviter les pièges de cette nature.

C'est pour tenter de prendre la mesure de cette catégorie d'anomalies que l'on a créé un fichier unique contenant une sélection d'informations individuelles issues des deux exploitations.

Le travail n'a été fait que pour le recensement de 1999. Il va de soi que l'on aurait pu effectuer la même démarche pour les recensements antérieurs. On aurait trouvé aussi probablement de graves erreurs (ainsi ce n'est pas la première fois que la codification des salariés de Peugeot-Mulhouse est erronée) Ce n'est pas la première fois non plus que les travailleurs frontaliers font l'objet d'une codification médiocre. S'agissant des frontaliers, nombre de mes collègues (et moi-même d'ailleurs) ont pourtant - depuis longtemps[7] - œuvré pour une prise en compte plus sérieuse des spécificités de cette catégorie d'emploi.

2 - La structure du fichier

Il s'agit d'un fichier individuel comprenant 56 millions d'enregistrements (seuls les ménages ordinaires sont pris en compte). Ce fichier comprend deux sous-ensembles :

° le sous-ensemble formé par les « *trois-quarts* », c'est-à-dire les personnes qui ne sont présentes que dans l'exploitation exhaustive.

° le sous-ensemble formé par le « *quart* ». Comme les personnes de ce groupe sont susceptibles d'être codifiées différemment dans les deux exploitations, on dédouble les variables (ainsi une localisation du travail selon l'exhaustif et une autre pour le quart).

[6] Le mal est encore accentué par le fait que l'on réalise des actions de diffusion plus grande que par le passé (naguère, les données de la statistique publique échappaient encore aux nécessités du marketing.

[7] Ainsi Nicole Séligman, dans la foulée de l'exploitation du recensement de 1975.

Les variables correspondant à l'exhaustif sont affectées du suffixe 1 (DCLT1 par exemple pour la commune de travail), celles du quart du suffice 4 (DCLT4).

Tableau 19 - Un exemple simple

		Commune de résidence	**Commune de travail**			
			L'exhaustif		**Le quart**	
	IND	DCR	DCLT1	SC1	DCLT4	SC4
les 3/4	E	**67482**	**67482**	1	-	-
	E	**67482**	67001	1	-	-
	E	67001	-	1	-	-
le 1/4	S	**67482**	**67482**	1	**67482**	4
	S	**67482**	67001	1	**67482**	4
	S	**67482**	**67482**	1	67001	4
	S	**67482**	67001	1	67001	4
	S	**67482**	67001	1	67002	4

(le code 67482 est celui de la commune de Strasbourg, les autres codes sont quelconques
DCR correspond à la commune de résidence, prise comme référence, DCLT à la commune de travail. Les pondérations affectées à chaque individu sont notées SC1 - toujours 1 par construction - et SC4 (en général 4, parfois 1)..

Les trois lignes du haut du tableau 1 correspondent à des personnes qui ne figurent que dans l'exhaustif (l'exploitation principale). Les personnes inactives ou en chômage n'ont évidemment pas de commune de travail. Les cases correspondantes des variables du sondage sont vides par construction.

Les lignes du bas du tableau 1 correspondent aux personnes appartenant au sondage (exploitation complémentaire). Elles correspondent à différents cas de figure, selon qu'il y a ou non cohérence entre les informations codifiées pour la commune de travail.

3 - Mode de création du fichier fusionné

Pour créer le fichier fusionné, on procède à des opérations de tri sur tous les ménages et les individus composant le ménage. Ces opérations (affectation d'un numéro d'ordre de la personne la plus âgée à la plus jeune en prenant en compte l'année de naissance, le mois et le quantième) sont identiques sur les deux fichiers.

Les deux fichiers élémentaires sont fusionnés. La partie « sondage » des variables de la population des « trois-quarts » est à blanc. Par convention on donne aux variables du quart des noms se terminant par le chiffre 4, et à celles de l'exhaustif des noms se terminant la valeur 1. Ainsi pour la commune de travail, on dispose d'une colonne pour l'exploitation principale (DCLT1) et une autre pour l'exploitation complémentaire (DCLT4). Les pondérations figurent

dans les colonnes SC1 (valeur 1 par construction) et SC4 (4 ou blanc selon les cas, 1 quand il s'agit des quelques communes où l'exploitation complémentaire concerne tous les individus).
Les données de base ayant été ainsi structurées, on peut tout d'abord aisément repérer les variables pour lesquelles il n'y a aucune divergence de codification. Ces variables, les plus nombreuses, sont d'ailleurs identiques par construction. C'est le cas du sexe, de l'âge… et des caractéristiques du logement.

En totalisant les effectifs des colonnes SC1 (valeurs toujours égales à 1) et SC4 (valeurs égales le plus souvent à 4, parfois à 1, lorsque l'exploitation complémentaire porte sur l'ensemble des logements d'une ville), on obtient les effectifs respectivement de l'exhaustif et du quart tels qu'on les obtiendrait en interrogeant les fichiers-détails.

On observe des divergences au moins dans les cas suivants :
- ° résidence antérieure pour les enfants nés pendant la période intercensitaire (codes DCRA1 et DCRA4)
- ° nature de l'emploi (EMPL1 et EMPL4)
- ° localisation du travail (DCLT1 et DCLT4).

Pour faciliter la compréhension des tableaux suivants, on introduit les symboles suivants :
QQ : indicateur de cohérence du lieu de travail
QT : indicateur de localisation du travail, par rapport à la résidence

Tableau 20 - les différents cas de figure

QQ	QT	Cas de figure
0	09	DCLT1 ne DCR
(les 3/4 seulement)	19	DCLT1 = DCR
1	00	DCLT1 ne DCR & DCLT4 ne DCR
(cohérence)	14	DCLT1 = DCR & DCLT4 = DCR
9	00	DCLT1 ne DCR & DCLT4 ne DCR
(incohérences)	04	DCLT1 ne DCR & DCLT4 = DCR
	10	DCLT1 = DCR & DCLT4 ne DCR

Les codes QT = 04, 10 et 00 correspondent aux anomalies