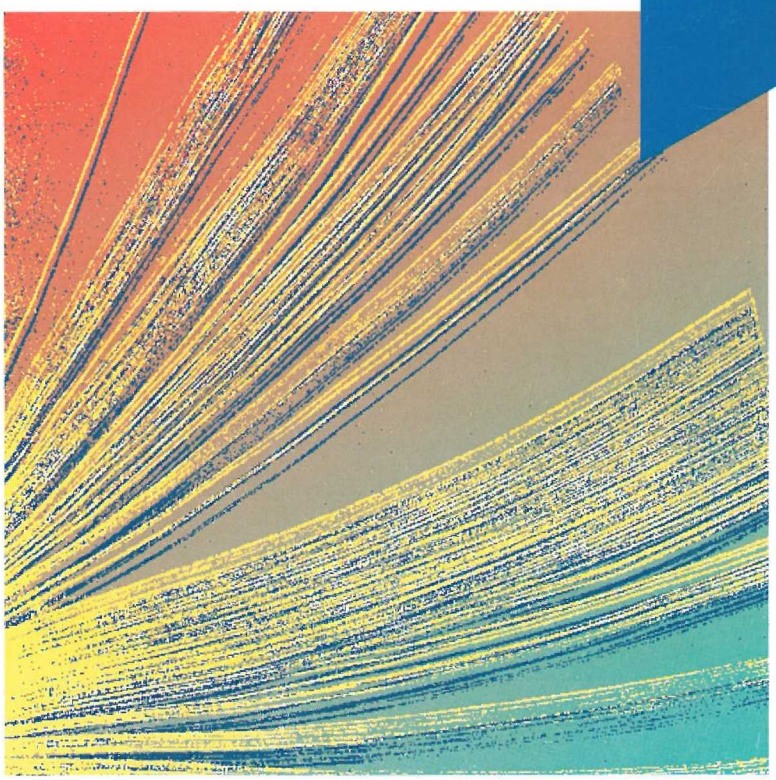
INSEE MÉTHODES
N° 69-70-71

INSEE MÉTHODES

ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

11 et 12 décembre 1996



INSEE

ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

11 et 12 décembre 1996

Les résultats contenus dans ce livre sont
le fruit d'un travail de recherche.
Ils n'engagent que leurs auteurs.

RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL
DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Direction Générale
18, boulevard Adolphe Pinard - 75675 Paris Cedex 14

Directeur de la publication : Paul Champsaur
La rédaction de ces actes a été coordonnée par Stéphane Tagnani
Maquettistes : Patricia Landais et Serge Fosseries

SOMMAIRE

SESSION 4 : LES STATISTIQUES LOCALES

SESSION 5 : LES ENQUÊTES SUR DES SUJETS SENSIBLES

PRÉSENTATION

Olivier Sautory
Chef de la Division "Méthodologie d'élaboration et d'analyse des données" de l'Insee

Les cinquièmes "Journées de méthodologie statistique", organisées par l'Unité "Méthodes statistiques" de l'Insee, se sont déroulées dans les locaux de l'ENSAE les 11 et 12 décembre 1996. Le succès de cette manifestation est toujours grandissant, puisque près de 600 personnes se sont inscrites, la plupart d'entre elles ayant assisté à plusieurs sessions.
Cette année, une attention particulière a été portée aux questions de méthodologie de collecte de données, avec notamment des séances consacrées aux questionnaires, aux enquêtes sur des sujets sensibles, et à l'expérience canadienne d'enquêtes par téléphone, présentée par un collègue de Statistique Canada. D'autres thèmes furent abordés au cours de ces journées : des développements récents dans le domaine des séries temporelles, les derniers travaux de l'Insee sur le thème des statistiques locales, l'usage des indicateurs d'inégalité des distributions de revenus. La conférence inaugurale a traité des rapports étroits et complexes entre statistique et sociologie, et un second collègue canadien a abordé la question de la confidentialité des données à l'aide des méthodes de "brouillage".

Dans la conférence inaugurale, **Alain Desrosieres** (Insee-Crest) a cherché à montrer la réciprocité des apports respectifs de la méthodologie statistique et de la sociologie. Dans la phase de construction des données, ces échanges peuvent porter sur la définition, la consistance et la fiabilité des variables, sur la structure et la signification des nomenclatures, sur les questions d'interaction entre enquêteurs et enquêtés, sur les difficultés du codage, ou même sur l'organisation du système statistique, administratif, universitaire ou privé. Dans la phase d'analyse des données, la coexistence ou la concurrence entre grandes familles de méthodologies peuvent elles-mêmes faire l'objet d'une analyse sociologique.

Conférences spéciales

Dans la première conférence spéciale, "Confidentialité des données ou l'art du brouillage clair", **Jean-René Boudreau** (Statistique Canada) a exposé comment son institut aborde le problème de la confidentialité des réponses dans les produits de diffusion appelés "fichiers de microdonnées". Pour assurer cette confidentialité, on peut être amené à modifier les réponses (c'est-à-dire introduire du bruit dans les données) de certains répondants pour réduire ou éliminer la possibilité d'identification nécessaire à toute divulgation. Cependant, ces modifications ne doivent pas invalider l'utilisation des données. D'où les questions : comment peut-on brouiller les données d'une manière "intelligible" ? Doit-on brouiller les données globalement (toutes les données) ou localement (seulement certaines données) ? Comment choisit-on les données qui seront brouillées ?

La pratique des enquêtes téléphoniques a, au cours des années, pris de plus en plus d'importance au sein du programme d'enquêtes de Statistique Canada. Plusieurs raisons peuvent expliquer ce fait. Cette méthode offre à la fois flexibilité, efficacité, et rapidité d'exécution, et peut, dans plusieurs situations, se substituer avantageusement aux interviews sur place très coûteuses. Par ailleurs, elle constitue un complément essentiel aux enquêtes postales, et se prête très bien aux méthodes assistées par ordinateur. **Jean-François Gosselin** (Statistique Canada) a présenté une vue d'ensemble de la pratique des enquêtes téléphoniques dans son institut, et a indiqué les questions et inquiétudes qui influeront grandement sur l'avenir de ce mode de collecte, concernant en particulier le taux de réponse aux enquêtes.

Les questionnaires et réponses aux enquêtes

La qualité d'une enquête est liée en particulier à la qualité du questionnaire utilisé pour la collecte. La conception et l'évaluation du questionnaire sont donc des étapes très importantes dans le développement d'une enquête. **France Bilocq** (Direction Générale de l'Insee (DG), unité méthodes statistiques, et Statistique Canada) a mis en évidence un certain nombre de principes de base destinés à minimiser les erreurs de réponse, en axant son intervention sur la formulation des questions et sur les différentes méthodes d'évaluation des questionnaires. La nécessité de la mise à profit de l'expérience des enquêteurs lors de cette phase de l'enquête a été soulignée.

Dans les enquêtes par sondage auprès des ménages ou des individus, de nombreuses questions peuvent être posées mais sur un échantillon de taille réduite. Au contraire, le caractère exhaustif d'un recensement de la population, s'il autorise des analyses sur des sous-populations très fines, limite l'ampleur et la complexité du questionnement. **Jacqueline Lacroix** (DG, division recensements de la population) a fait le point sur l'élaboration des questionnaires du 33ème recensement de la

population. Elle a présenté en particulier les résultats de tests de collecte permettant de juger de l'acceptabilité et de la compréhension de nouvelles questions, retenues après une large consultation, menée dans le cadre du CNIS auprès des principaux utilisateurs des données.

Depuis 1993, plusieurs enquêtes auprès des ménages réalisées par l'Insee se sont vues refuser par la CNIL le statut obligatoire, jusqu'alors systématiquement accordé. Afin de mesurer l'impact de l'obligation de réponse sur le taux de réponse à ses enquêtes ménages, l'Insee a greffé un dispositif d'observation sur l'enquête "permanente conditions de vie des ménages" de janvier 1996 : deux échantillons de 2000 ménages ont été enquêtés dans des conditions très proches, la seule différence tenant à l'annonce du statut de l'enquête. **Catherine Berthier** et **Françoise Dupont** DG, unité méthodes statistiques) ont présenté les premiers résultats de cette opération, dont le principal constat est le suivant : le comportement de refus des enquêtés au premier contact est nettement plus fréquent en l'absence de l'obligation (18,2% contre 8,6%), et la tentative de relance aboutit moins souvent.

Les séries temporelles

Deux nouveaux logiciels de désaisonnalisation sont apparus récemment : X12-ARIMA et TRAMO-SEATS. **Ketty Attal** (DG, unité méthodes statistiques) a consacré la première partie de son exposé à une présentation générale des méthodes de désaisonnalisation, du point de vue historique et du point de vue de leur philosophie. Elle a ensuite décrit les deux nouveaux logiciels, qui reposent sur des principes très différents, mais qui ont en commun d'une part l'utilisation des modèles ARIMA, d'autre part le souci d'améliorer la "qualité" de la série qui doit être décomposée, par un traitement préalable assez poussé.

Christophe Planas (Eurostat) s'est ensuite attaché à la comparaison de ces deux logiciels d'un point de vue théorique. Il a présenté les propriétés des différents filtres intervenant dans les logiciels, et les conséquences pratiques ont pu être vérifiées sur la série de la production industrielle mensuelle en France.

Les enquêtes de conjoncture réalisées par l'Insee offrent un type d'information particulièrement utile pour l'analyse conjoncturelle. Cependant, le nombre et la diversité des questions posées rendent souvent délicate l'interprétation des résultats obtenus. La communication de **Catherine Doz** (Direction de la Prévision) et **Fabrice Lenglart** (DG, division synthèse conjoncturelle) a présenté la construction d'un indice composite susceptible de résumer cette information, qui utilise les modèles à facteurs dynamiques. Les méthodes standard d'analyse factorielle classique ont également été utilisées, conduisant à des résultats très proches des précédents, sur les données des enquêtes mensuelle et trimestrielle de conjoncture dans l'industrie.

Les mesures d'inégalité

Dans un exposé introductif, **Olivier Sautory** (DG, unité méthodes statistiques) a présenté les principaux indicateurs utilisés pour mesurer l'inégalité (dans le cas de distributions de revenus en particulier), ainsi que leur interprétation et leurs propriétés : signification statistique, valeurs minimale et maximale, compatibilité avec l'ordre (partiel) des courbes de Lorenz, sensibilité à des transferts entre individus, décomposabilité additive.

Parmi ces mesures d'inégalité, le coefficient de Gini est un des plus utilisés pour évaluer les disparités de revenus. De par sa forme fonctionnelle "fortement" non linéaire, il apparaît a priori difficile d'évaluer la précision d'un tel indice quand il est calculé sur des données obtenues dans une enquête par sondage, car la théorie habituelle ne s'applique pas. **Jean-Claude Deville** (DG, unité méthodes statistiques) s'est attaqué à ce problème, et a montré que l'estimation de la variance du coefficient de Gini se ramène à l'estimation de la variance du total d'une variable artificielle linéarisée assez exotique mais facile à construire.

Jérôme Accardo et **Madior Fall** (DG, division revenus et patrimoine des ménages) ont comparé les indicateurs d'inégalité des revenus tirés de différentes enquêtes-ménages réalisées par l'Insee de 1984 à 1993. Utilisant la méthode proposée par J.-C. Deville, ils ont calculé la précision des coefficients de Gini, et ont conclu à l'absence de significativité des évolutions de cet indicateur d'une enquête à l'autre. Ils ont également étudié la robustesse de l'indicateur, en évaluant l'impact de "contaminations" simulant différents types d'erreur de mesure : ces simulations tendent à prouver l'instabilité de l'indicateur, résultats que vient nuancer cependant, selon les auteurs, le constat de la faible dispersion du coefficient de Gini dans les diverses enquêtes, qui reposent pourtant sur des modes de collecte assez différents.

Les statistiques locales

La réponse à la demande de résultats locaux - de plus en plus forte depuis quelques années - peut s'obtenir par deux méthodes : en réalisant une enquête locale, ou en exploitant une enquête nationale par des techniques adaptées. **Sophie Destandau** (DG, unité méthodes statistiques) a dressé un panorama des méthodes d'estimation sur petits domaines à partir d'une enquête par sondage, en précisant les avantages et les inconvénients des différentes méthodes d'estimation : estimations "directes", estimations "synthétiques", estimations "combinées". A titre d'illustration, certaines de ces techniques ont été appliquées pour obtenir des résultats régionaux sur l'enquête Education de 1992, et les performances de ces estimateurs ont pu être comparées.

La France ne disposant pas de registres de population, les recensements de la population y constituent la base du système d'informations socio-démographiques. Cependant, entre deux recensements, l'actualisation de certaines données est nécessaire, notamment à un niveau géographique fin, d'autant plus que les recensements ont tendance à s'espacer. Un groupe de travail a été mis en place à l'Insee, avec pour objectif de proposer une nouvelle méthode d'estimations locales de population. **Jean-Claude Labat** (DG, direction des statistiques démographiques et sociales) a présenté les principales conclusions de ce groupe : réaliser une synthèse efficace et robuste des informations apportées par différentes sources administratives et mobiliser un nombre suffisant de "bonnes" sources, constituant ainsi un système "multi-sources" souple et fiable, sans être trop complexe.

L'Insee a élaboré une nouvelle nomenclature spatiale, le "Zonage en aires urbaines" : elle remplace les Zones de peuplement industriel ou urbain (ZPIU) qui, définies en 1962, couvraient en 1990 les trois quarts du territoire et abritaient 96 % de la population. **Thomas Le Jeannic** (DG, division statistiques et études régionales) a montré en quoi ce nouveau zonage constitue une nouvelle approche territoriale de la ville et de son espace périurbain, dont il donne une image plus restrictive : défini cette fois autour de 361 pôles urbains, sélection des plus grandes unités urbaines, le périurbain doit en outre dépasser un seuil élevé de migrants alternants (40%) vers ces pôles.

Les enquêtes sur des sujets sensibles

Alain Giami et **Alfred Spira** (Inserm) ont présenté un bilan méthodologique de trois enquêtes françaises sur le comportement sexuel : l'enquête ACSF auprès des adultes métropolitains, l'enquête ACSAG auprès des adultes en Guadeloupe, Martinique et Guyane, et l'enquête ACSJ auprès des jeunes de 15 à 18 ans en métropole. La méthodologie de ces enquêtes est bien différenciée : téléphone pour l'enquête ACSF, face à face au domicile des adultes pour l'enquête ACSAG , face à face dans les établissements scolaires ou d'apprentissage pour l'enquête ACSJ. Les méthodes d'échantillonnage employées, toujours probabilistes, s'avèrent également d'une grande diversité (surreprésentation des situations à risque vis-à-vis du sida dans l'échantillon ACSF, grâce à une vaste enquête-filtre). Les intervenant ont également mis l'accent sur les difficultés soulevées par de telles enquêtes : comment interroger quelqu'un sur sa sexualité, en respectant sa sensibilité.... en particulier pour les jeunes ? Quels mots employer ? Quel sens donner aux divergences de réponse des hommes et des femmes sur leurs pratiques conjointes ? Une telle recherche a nécessité le travail en commun d'une vaste équipe pluridisciplinaire formée de psychologues, psychosociologues, linguistes, sociologues, économistes, médecins épidémiologistes, démographes et statisticiens !

L'Ined et l'Association l'Elan Retrouvé ont réalisé pendant les hivers 1994-1995 et 1995-1996 deux enquêtes auprès d'un échantillon représentatif de personnes sans domicile utilisatrices de services d'hébergement et de distribution de nourriture, à Paris. **Maryse Marpsat** et **Jean-Marie Firdion** (Ined), et **Viviane Kovess** (Association l'Elan Retrouvé) ont exposé la méthodologie largement commune de ces enquêtes, inspirée des méthodes de sondage et d'estimation d'enquêtes américaines. Il se pose bien sûr des problèmes d'échantillonnage, puisque la population échappe à la base de sondage classique des logements, mais aussi des questions de fond comme d'éthique, ainsi que les difficultés pratiques que l'on peut imaginer. L'enquête de l'Ined, outre sa fonction d'expérimentation pour une enquête à plus grande échelle, cherchait à explorer les processus conduisant à la situation de sans-domicile. L'enquête de l'Elan Retrouvé avait pour objectif de connaître la prévalence des principaux problèmes de santé mentale dans la population des sans-abri, la présence de maladies somatiques sévères et de handicaps, et l'utilisation de soins.

Conférence inaugurale

LES APPORTS MUTUELS DE LA MÉTHODOLOGIE STATISTIQUE ET DE LA SOCIOLOGIE

Alain Desrosieres

L'expression même de "méthodologie statistique" implique une division du travail entre, d'une part, des "experts" de l'outil statistique en tant que tel et, d'autre part, des "usagers" de celui-ci, économistes, sociologues, historiens ou psychologues. Dans cette perspective, les progrès et les innovations de la statistique auraient un impact sur chacune de ces sciences, perçues comme des domaines d'application de formalisations élaborées en dehors d'elles. Il est vrai que l'autonomisation croissante de la recherche et de l'enseignement des statistiques mathématiques tend à conforter cette impression. Pourtant l'histoire et la sociologie de la statistique offrent de nombreux exemples d'effets *de sens inverse*, où des questions posées par des sociologues, des économistes ou des spécialistes d'autres disciplines ont profondément transformé la méthodologie statistique. Un exemple important en est fourni, à la fin du XIXe siècle, par les travaux de Francis Galton et Karl Pearson sur l'hérédité humaine et l'eugénisme, qui ont conduit à formaliser la régression linéaire, la corrélation et le test du chi-deux (Stigler, 1986). A leur suite, les recherches agronomiques de Ronald Fisher conduisent à la statistique inférentielle moderne et à l'analyse de la variance.

Ce sont bien sûr les développements de l'économétrie qui, depuis les années 1930 et 1940, offrent les exemples les plus connus des interactions réciproques entre une discipline scientifique et la théorie statistique. C'est même à l'occasion des controverses de cette période cruciale de la quantification des sciences sociales, qu'ont été distinguées trois façons différentes d'envisager les relations entre "théorie" et "données observées" (Morgan, 1990). Pour la première, une théorie, de type hypothético-déductif (Walras) peut, au mieux, être illustrée par des données, mais elle est logiquement supérieure à celles-ci. Pour la seconde en revanche, l'induction permet d'inférer, à partir de régularités statistiques observées, des "lois" générales (Quetelet), dont la nature est complètement différente des "lois" postulées par les théories déductives du premier type. Enfin, dans une troisième perspective, qui n'apparait que dans les années 1930, des hypothèses théoriques sont *testées*, rejetées ou provisoirement acceptées, au moyen des outils de la statistique inférentielle alors naissante. Ces trois manières d'articuler "théorie" et "données" sont clairement visibles dans l'histoire de la pensée économique, où, jusqu'aux années 1940, se sont affrontés, d'une part, des économistes logiciens et mathématiciens, et, d'autre part, des économistes dits "historicistes", ou "institutionnalistes", *grands utilisateurs de*

statistiques descriptives, mais hostiles aux "lois" théoriques générales. L'approche probabiliste de l'économétrie (Haavelmo, 1944) transforme profondément ce paysage manichéen, en introduisant la notion de "modèle stochastique".

Cette distinction nette entre trois façons de relier la théorie et les données permet de reconstruire l'histoire longue des relations entre économie et statistique : Mary Morgan (1990) ou Michel Armatte (1995) offrent de bonnes synthèses de cette histoire. Mais, dans cette histoire, le pôle "théorie" est relativement dominant. Depuis Adam Smith, les controverses entre économistes ont plus souvent porté sur le corpus plus ou moins formalisé et cohérent des hypothèses théoriques que sur l'observation et la construction des données, laissées à des spécialistes souvent extérieurs à la discipline : les statisticiens, les comptables nationaux. La dernière période, où des modèles sont proposés et testés au vu de données empiriques n'a que partiellement modifié cette situation, même si, dans le cas français de l'Insee qui rassemble statisticiens et économistes, cette séparation est moins nette que dans les pays anglo-saxons.

Le cas de la sociologie est très différent. Celle-ci n'a jamais disposé d'un corps théorique autour duquel se noueraient les controverses. Par contre, le pôle "statistique" y a joué, historiquement, un rôle essentiel. Parmi les sciences humaines, la sociologie se distingue par son hypothèse centrale : une société doit être étudiée en soi, et non comme la simple juxtaposition des individus qui la composent. Dans ses versions dites "holistes", ou "réalistes", cette hypothèse est forte : les propriétés du groupe sont radicalement distinctes de celles des individus. Le groupe a une réalité par lui-même. En revanche, dans des versions plus "individualistes", ou "nominalistes", l'individu reste premier, mais des effets de composition, d'interaction (tels que les "effets pervers"), d'imitation ou de panique, suffisent à donner à la sociologie un statut distinct et spécifique. Mais, qu'ils tiennent pour l'une ou l'autre de ces deux versions, les sociologues ont traditionnellement cherché à explorer cette hypothèse en suivant deux voies (non indépendantes) : la statistique et le droit. La statistique est du social observé *in vivo*, le droit est du social cristallisé, durci, déposé dans des institutions (Héran, 1984). Mais, comme les statisticiens et les sociologues quantitativistes ne l'observent pas toujours, la statistique est souvent tributaire du droit et des pratiques administratives, comme dans le cas des statistiques du crime et du suicide, chères à Quetelet et à Durkheim. Le durcissement institutionnel est même une condition *sine qua non* de la "fiabilité" des statistiques : les polémiques récurrentes sur la mesure du chômage le démontrent tous les jours. Le développement du réseau statistique est lié à celui d'un système d'institutions. Cet investissement, analogue à celui d'un réseau routier ou ferroviaire, crée des catégories qui deviennent ensuite incontournables, le plus souvent différentes d'un pays à l'autre. De tout cela, le sociologue moderne qui recourt à d'inépuisables "banques de données" est quelquefois peu conscient.

Il est vrai que, dès lors qu'elle est coupée de ses conditions sociales et administratives d'enregistrement, la statistique offre, par ses régularités de mieux en mieux explorées par les développements méthodologiques, des arguments "clé en main" pour justifier l'autonomie et la prééminence d'une science nettement distincte de la philosophie, du droit, de la psychologie, ou même de l'économie. Le groupe social a des structures spécifiques et des propriétés de régularité et de prévisibilité, dont sont dénués les individus, volatiles et imprévisibles. Cette version classique de la sociologie quantitative a été clairement formulée par Quetelet, dans les années 1830, puis approfondie par Durkheim, Lazarsfeld, Bourdieu, l'Insee, l'Ined et beaucoup d'autres. Elle a été critiquée de plusieurs façons. Des sociologues inspirés par la théorie économique ont cherché à réintégrer des hypothèses de rationalité des agents individuels. Identifiant la sociologie à l'analyse des effets de composition ou d'agrégation de ces comportements rationnels, ils la rapprochent ainsi de la démarche classique des économistes qui partent de la théorie.

Les apports de la critique ethno-méthodologique

D'autres sociologues, d'inspiration toute différente de celles des deux catégories précédentes, ont procédé, à partir des années 1960, à ce qu'ils percevaient alors comme une critique radicale de la sociologie quantitative, et particulièrement de la sociologie d'enquête par questionnaire, qui s'était répandue rapidement aux Etats-Unis, notamment sous l'influence de Lazarsfeld. Ils attaquaient la "méthodologie", selon eux réductrice et plaquée de l'extérieur, des sociologues statisticiens. Ils lui opposaient une "ethno-méthodologie"[1], c'est-à-dire une analyse des façons dont les personnes enquêtées elles-mêmes, comprennent, décrivent et catégorisent leurs propres activités (Cicourel, 1964). La comparaison entre les schèmes de perception des acteurs et ceux des sociologues tournait souvent au désastre pour ces derniers, et ceci d'autant plus que l'écart social et culturel entre le premiers et les seconds est plus grand. Cette critique, qui se voulait ravageuse de la sociologie quantitative, a été ensuite intégrée en partie dans la méthodologie de fabrication et d'interprétation des questionnaires, notamment en matière d'opinion. Une attention plus grande a été portée aux questions de vocabulaire, et, plus profondément, de signification des questions, différente selon les milieux sociaux. Ainsi par exemple, la question, posée à des femmes : "Etes vous favorable au travail des femmes ?" était encore, dans les années 1960 et 1970, entendue très différemment, par les femmes de milieux ouvrier et cadre. Les premières percevaient ce travail comme un pis-aller, inévitable si le salaire du mari était insuffisant. Les secondes, en revanche, y voyaient le signe de l'émancipation de la femme, et y étaient favorables.

---

1 On pourrait dire aussi : "méthodologie indigène".

Dans le climat américain des années 1960 et 1970, les critiques des ethno-méthodologues contre les sociologues statisticiens reflétaient un clivage profond dans le groupe des sociologues, perceptible même géographiquement. Les premiers étaient plutôt sur la côte Ouest, et les seconds sur la côte Est. Il était donc peu probable que de réels échanges aient lieu entre eux. Ainsi le riche article de Clifford Clogg (1992), publié dans *Statistical Science*, sur l'"impact de la méthodologie sociologique sur la méthodologie statistique", ne mentionne même pas l'existence du livre de Cicourel (1964) : "*Methods and measurement in sociology*". Cet article de Clogg explique en détail comment des innovations concernant les méthodes de *traitement* et *d'analyse* statistique des données ont été induites par des questions sociologiques, mais il n'insiste pas du tout sur la *construction* de ces "données". Le partage entre les deux "méthodologies" semble complet.

En France, du fait de l'existence d'institutions comme l'Insee et l'Ined, qui concentrent la conception, la réalisation, l'analyse et l'interprétation des enquêtes, ce clivage est moins marqué qu'en Amérique. Dès les années 1950, les enseignements, à l'école de l'Insee, de Chevry, Croze et Desabie contiennent de nombreuses indications pratiques manifestant une grande sensibilité aux dimensions sociologiques des formulations des questions, de l'interaction entre enquêteurs et enquêtés, et des problèmes de codage des cas ambigus, trois domaines de prédilection des critiques des ethno-méthodologues américains. A partir des années 1960, des sociologues universitaires, Bourdieu, Baudelot, Héran et d'autres, enseignent à l'ENSAE, et contribuent à construire théoriquement les conseils pratiques de la première génération des statisticiens-enquêteurs de l'Insee.

Ainsi, les définitions des variables et des catégories, auparavant perçues souvent comme des questions de logique formelle et de bonne organisation des instructions adressées aux ateliers de chiffrement, sont réinterrogées dans une perspective sociologique et historique. Ces définitions sont le plus souvent des *conventions d'équivalence* entre des cas hétérogènes. L'origine de ces conventions peut être externe et antérieur au travail du statisticien, par exemple inscrite dans du droit ou des coutumes. Elle peut aussi être produite par le statisticien lui-même au moment de la conception ou même de l'exploitation de l'enquête. La distinction de ces deux cas est essentielle pour la phase ultérieure *d'interprétation* des résultats obtenus. Ainsi, l'étude des nomenclatures, auparavant perçue comme austère et ingrate, devient, dans le contexte particulier de la France des années 1970 et 1980, un fécond thème de recherche sociologique et historique : sur les postes de dépense des enquêtes budgets de famille (Boltanski, 1970), sur les branches industrielles (Guibert, Laganier, Volle, 1971), sur les catégories socioprofessionnelles (Desrosières, 1976 ; Boltanski, Thévenot, 1983 ; Merllié 1990 ; Kramarz, 1991), sur le chômage (Salais, Baverez, Reynaud, 1986), sur les causes de décès (Fagot-Largeault, 1989), sur la criminalité (Robert *et alii,* 1994). Dans toutes ces recherches, deux questions distinctes bien que très liées sont soulevées : celle de la définition théorique et pratique des classes, et celle du codage, c'est-à-dire du travail

concret d'affectation d'un cas à une classe. Un des apports de ces recherches est précisément de montrer que la première phase (définition des classes) ne peut jamais être pensée indépendamment de la seconde. Celle-ci est, en définitive, une des plus suggestives du travail du sociologue statisticien, ce qu'ignore en général l'épistémologue théoricien et logicien : les "critères" les plus rigoureux sont souvent balayés par l'exploration d'une pile de questionnaires. Les critiques des ethno-méthodologues sont dès lors réintégrées dans la pratique du statisticien.

Soit, par exemple, la question extrêmement sensible, socialement et politiquement, du partage de la population totale en trois catégories : population active occupée, population active sans emploi (ou chômeurs), population inactive. Les frontières entre ces trois groupes peuvent être définies par des critères généraux (variables d'un pays à l'autre), ou avec le Bureau international du travail (BIT), qui est supposé fixer une norme internationale en la matière. Mais l'application concrète de ces règles générales multiplie les difficultés et les hésitations. Cette situation s'aggrave même en période de crise de l'emploi, où les critères juridiques sont de plus en plus tournés par des situations dites "informelles", qui contraignent le statisticien à hésiter entre "le droit" et "le fait". Ces situations et ces ambiguités mettent mal à l'aise le statisticien, sommé par la presse et les usagers de donner "le bon chiffre", approchant au mieux "la réalité". Il s'en tire souvent par des expressions comme "halo" ou "flou", comparable à la situation de l'astronome visant une étoile avec un instrument mal réglé, ou à celle du myope qui aurait perdu ses lunettes.

La sociologie quantitative et ses trois modèles de réalité

Cette pression à fournir le "bon chiffre", comme les métaphores sur le flou, résultent de ce que la statistique sociale a été construite, légitimée et diffusée à partir du modèle métrologique réaliste des sciences de la nature. La réalité existe antérieurement à son observation, comme l'étoile polaire a existé bien avant tous les astronomes. Mais précisément la définition et la mesure de la population active et du chômage relèvent d'une autre épistémologie que celle de l'étoile polaire. Elles impliquent des *conventions* (analogues aux principes généraux des lois et des codes votés par les Parlements) et des *décisions* (analogues à celles d'un juge) d'affecter tel cas à telle classe. Pour certains domaines, comme la statistique criminelle, cela semble presque évident, bien que, même dans ce cas, la demande "réaliste" surgisse toujours.

Les questions soulevées ici relèvent de la sociologie, mais non pas au sens de la sociologie quantitative usuelle, qui utilise et interpète sociologiquement des données *auparavant* construites, mais au sens d'une sociologie réflexive, qui étudie les usages sociaux et les rhétoriques d'interprétation de ces données statistiques. Une des raisons pour lesquelles la sociologie quantitative a toujours eu des difficultés à trouver sa place dans une discipline sociologique (de toute façon très émiettée en

paradigmes différents), tient à son hésitation sur le statut de réalité des objets "mesurés". On peut, schématiquement, distinguer trois façons distinctes d'interpréter ces données. Elles sont inspirées par trois grandes familles de disciplines scientifiques : les sciences de la nature, les sciences de la vie, les sciences juridiques et politiques. Chacune implique une conception différente de "la réalité", entre lesquelles la sociologie oscille, non pas pour des raisons épistémologiques, mais en fonction des *contraintes spécifiques à ses divers usages.* Une étude sociologique des usages de la statistique, notamment en sociologie, implique une explicitation approfondie de ces contraintes de situation : qui va lire ? avec quelles notions *a priori* des objets manipulés et du réalisme de leur mesure ? pour faire quoi (argumenter, contester, décider, etc...) ?

Les sciences de la nature (astronomie, physique), ont imposé, dès le XVIIIe siècle une épistémologie de la mesure, enserrée par des schèmes probabilistes. La "loi des erreurs" a inspiré Quetelet pour construire son "homme moyen". Cette métrologie réaliste s'est imposée comme un modèle premier, dont les statisticiens ne peuvent jamais se défaire complètement, ne serait ce que parce qu'elle leur est inlassablement demandée : "quel est le *vrai* chiffre du chômage ? de la hausse des prix ?". Les commentaires d'accompagnement sur le caractère *conventionnel* de ces mesures ne peuvent rien contre une demande sociale, entretenue aussi en partie par les statisticiens eux-mêmes, soucieux de s'inspirer du modèle le plus achevé, celui des sciences de la nature. Le succès du mot "mesure", attesté par une exposition récente et par des ouvrages nombreux, est un signe de la prédominance implicite de ce modèle, là ou d'autres mots, comme, d'une part, "indice" ou "symptôme", ou, d'autre part, "convention d'équivalence" ou "domaine d'action", typiques des deux autres modèles, pourraient être utilisés.

Le langage des *sciences de la vie,* fort différent du précédent a été utilisé très tôt par Quetelet et par les hygiénistes, adeptes de la "statistique morale". Les "moyennes subjectives", les "propensions" au mariage, au crime ou au suicide, calculées à partir de statistiques administratives globales, étaient des indicateurs macrosociaux, révélés par les régularités statistiques et supposés consistants, *reflétant* des attributs de la société impossibles à atteindre directement. Plus tard, les "variables latentes" de la sociologie anglo-saxonne auront les mêmes propriétés que les "propensions" de Quetelet. Impossibles à mesurer directement, elles apparaissent comme des résultats plus ou moins robustes de l'analyse statistique des variables "patentes", et sont supposées refléter un contenu sociologique plus profond et plus généralisable que ces dernières. Les axes factoriels de l'analyse des données à la française ont des propriétés comparables. L'"intelligence générale" ou le "quotient intellectuel" de Binet et Simon en étaient des ancêtres. De même, un indice de prix, un indice de production industrielle ou un indice boursier (Dow Jones, CAC 40 ou Nikkei) sont à la fois des moyennes pondérées et des "variables latentes", plus générales et explicatives que chacunes de leurs composantes.

Ce langage de la variable latente est très différent de celui de la métrologie des sciences de la nature, même si les progrès de la métrologie moderne font de plus en plus apparaître les définitions des unités de longueur, de poids et de temps, comme des *conventions,* historiquement variables. La rhétorique réaliste *directe* reste de règle dans la plus grande partie des sciences de la nature (à l'exception peut être de la physique relativiste et de la mécanique quantique) et sert de modèle à une statistique sociale qui, par ailleurs, recourt *aussi* au langage des sciences de la vie, avec le réalisme *indirect* de ses indicateurs et de ses symptômes. Une partie des débats politiques et scientifiques autour de la statistique sociale résulte de ce flottement entre deux formes de réalisme, direct ou indirect, centrées sur l'idée de *mesure.* Le référent de celle-ci est visible, caché ou postulé par l'opération même de construction de la variable latente.

Mais ces deux formes de réalisme se distinguent d'une troisième rhétorique, conventionnaliste, où la trace de l'acte initial de codage reste visible et importante, soit dans une perspective de dénonciation, soit parce que l'agrégat est directement articulé sur une forme d'action collective. Dans ces cas, l'*intention* de l'agrégation et de l'addition reste présente dans l'usage de la statistique présentée. Soit, par exemple, une statistique récente et largement commentée, celle de la maltraitance à enfant, problème social et politique aujourd'hui jugé essentiel. Le fait qu'une question devienne "socialement jugée sociale", c'est-à-dire relevant d'une action publique, transforme son statut statistique. Des procédures de repérage (numéros verts), d'enregistrement et de comptage sont mis en place. Des définitions et des critères sont formulés. Quand cette opération est encore récente, les interprètes hésitent entre deux lectures : "le nombre des enfants maltraités a augmenté", ou "les procédures d'observation se sont améliorées". Ce flottement avait déjà été observé, à propos du chômage, dans les années 1960, avec la progressive mise en place de l'ANPE. Il met mal à l'aise les commentateurs, qui ne peuvent se résoudre à renoncer à une rhétorique réaliste, et critiquent les incertitudes du système d'observation qui ne peut leur fournir des "chiffres fiables". Mais c'est précisément parce que la question est devenue telle qu'on en parle dans les journaux, que ce système d'évaluation évolue vite, et est jugé "peu fiable". La "fiabilité" est étroitement associée à la stabilité et à la routinisation de la chaîne d'enregistrement et de comptage, qui impliquent que le sujet est devenu moins brûlant.

Mesures, indices ou classes d'équivalence

La sociologie quantitative est hantée, depuis Quetelet, par trois modèles, issus de trois types de sciences, mais elle l'ignore en général. Ces modèles impliquent des outils et des usages rhétoriques différents. Les *mesures* des sciences de la nature, comme les *indices* des sciences de la vie sont exprimés par des *variables continues* directement observables dans le premier cas, et "latentes" dans le second cas. Mais ces variables caractérisent tout l'univers étudié de façon uniforme. Elles peuvent être

modélisées, ajustées à des lois de probabilité, comparées, corrélées, regressées, testées selon les méthodes de la statistique inférentielle. Le modèle initial de ces méthodes a été fourni par les recherches de Ronald Fisher, au laboratoire d'agronomie expérimentale de Rothamstead, dans les années 1920 et 1930. Les "variables" sont des caractérisations homogènes interchangeables : la nature suit des lois générales et transposables d'une expérience à l'autre. Cette logique de la mesure a souvent été transférée telle quelle aux sociétés humaines par les premiers pionniers de la sociologie quantitative, notamment anglo-saxons. Par exemple, leur échelle sociale, continue et unidimensionnelle (utilisée pour étudier la mobilité sociale) est issue des travaux de Galton sur l'échelle des aptitudes *(abilities)*, reprise aussi par le psychologue Spearman (1904) pour étalonner une échelle d'intelligence générale (Gould, 1983). La recherche des "variables latentes", statistiquement plus efficaces et interprétables de façon synthétique, à la façon d'une moyenne, est au coeur de cette démarche, où les sciences de la vie cherchent à coller au plus près des sciences de la nature. Dans cette perspective, où les variables interagissent de façon uniforme, on peut étudier l'"effet d'une variable", comme l'expérimentateur agricole ajoute ou retire une quantité d'engrais.

Mais, dans le cas des sciences sociales, une troisième forme d'identification des objets de la statistique intervient. La *classe d'équivalence* est une *convention,* issue des sciences juridiques et politiques. C'est uns construction humaine, affectant des droits et des devoirs communs à une classe d'hommes définis par des lois, des règlements, des conventions ou de simples usages : "les hommes naissent et demeurent libres et égaux en droits". La société elle-même est une monumentale entreprise de taxinomie, dont la statistique enregistre les effets : le sexe, l'âge, le lieu de naissance, le diplôme, la catégorie socioprofessionnelle, le lieu de résidence, le statut matrimonial et familial, etc... Galton avait tenté de réduire le statut social à une échelle naturelle d'aptitudes innées, mais cet essai de naturalisation a fait long feu. La notion de "classe d'équivalence" est moins familière au statisticien, spontanément influencé par le modèle réaliste des sciences de la nature, où on procède à des *mesures*, et non à des opérations de *jugement*, visant à coder, c'est-à-dire à affecter, selon des conventions générales fixées *a priori* , des cas singuliers à des classes. Bien que les contraintes cognitives de ces opérations de jugement ne fassent pas partie de la culture du statisticien, celui-ci procède intensivement à ces opérations. Les discussions sur le "flou" ou le halo" de tel ou tel agrégat statistique reflètent la complexité de ces contraintes, et surtout la tentation de les rabattre sur le modèle réaliste des sciences naturelles.

La sociologie quantitative est donc vouée à manipuler en même temps plusieurs outillages, de statuts sociaux et techniques très différents, la *mesure* (sous sa forme directe ou "latente"), et la *classe d'équivalence*. Dans leur majorité, les sociologues anglo-saxons, proches historiquement des sciences expérimentales et de leurs outils, cherchent spontanément à construire des variables continues et mesurables. Celles-ci sont en petit nombre afin de pouvoir être confrontées, corrélées et testées dans des

modèles probabilistes, issus précisément des sciences expérimentales. Beaucoup de sociologues français, en revanche, peut-être de culture plus historique et philosophique, ont résisté à cette idée de mesure continue. Ainsi par exemple, les "échelles de prestige social", construites à partir d'enquêtes d'opinion sur les métiers, ont eu grand succès en Amérique et en Grande-Bretagne mais sont à peu près inconnues en France. En revanche, la nomenclature des catégories socioprofessionnelles a été particulièrement travaillée et utilisée en France, depuis Jean Porte dans les années 1950[2].

Presque dès son origine, la méthodologie statistique a été marquée par une controverse, entre Karl Pearson et son élève Udny Yule, entre 1900 et 1914, au point que leurs relations en furent affectées (Mac Kenzie, 1981). Le débat portait précisément sur la tension entre les notions de mesure et de classe d'équivalence: comment mesurer la forme de l'association, ou "corrélation" entre deux variables, quand celles-ci ne sont des mesures sur une échelle continue, mais des classements, ou "variables discrètes" ? Le cas le plus simple est un tableau de 2x2=4 cases, croisant deux tris dichotomiques. De tels cas sont usuels dans les sciences sociales, domaine oùYule tente de transférer les méthodes imaginées par Pearson pour les sciences biologiques.Yule cherche notamment à étudier les effets sur la pauvreté de deux modes d'assistance (à domicile ou en asile). Il propose un indicateur de "corrélation" facile à calculer. Si *a, b, c* et *d* dont les quatre cases du tableau, la force de l'association est mesurée par $Q=(ad-bc)/(ad+bc)$. Aucune hypothèse n'est nécessaire sur la distribution des 4 variables.

Karl Pearson, en revanche travaille, dans le cadre de la biométrie, sur des mesures continues, physiques ou non, dont la distribution typique suit une loi normale. Si les variables ne sont pas continues mais discrètes, son réflexe est de les rendre continues, en utilisant les fréquences observées des catégories pour étalonner celles-ci sur une échelle continue, *en supposant la distribution normale*. Ce tour de passe-passe lui permet de revenir à la situation continue, en bâtissant une loi normale à deux dimensions dont les distributions marginales s'ajustent sur les deux distributions marginales observées dans le tableau initial. Il démontre qu'il y en a une et une seule, et l'un de ses paramètres fournit la corrélation souhaitée, baptisée "coefficient de corrélation tétrachorique". Pearson n'a en revanche que dédain et sarcasmes pour la formule simple de Yule, $Q=(ad-bc)/(ad+bc)$. Elle est arbitraire et ne repose sur rien. Elle pourrait aussi bien être remplacée par Q^3 ou Q^5. La polémique, qui dure de longues années, porte sur les hypothèses de continuité, naturelles pour le biométricien, mais peu évidentes en sciences humaines. A un

2 Cette comparaison franco-anglaise sur les usages respectifs des *classes* et des *mesures* dans les deux sociologies a été étudiée en détail, à partir d'enquêtes anglaises et françaises, par Marie-Ange Schiltz (1991), dans une étude reprise, sous forme modifiée, dans Greenacre et Blasius (eds) 1994.

moment, Yule évoque une situation où des individus sont vivants ou morts. Que signifie donc, dans ce cas, une hypothèse de continuité ?

L'hybridation des outils et l'oubli de leurs origines

La recherche des généalogies et des implications sémantiques des deux notions de *mesure* et de *classe* ne doit pas laisser croire que, ensuite, elles ont mené des vies séparées. Toute la dynamique de la méthodologie statistique a poussé à les associer, les combiner, et à façonner des opérateurs d'échange entre elles. On en citera ici deux exemples, portant sur deux méthodes statistiques, souvent présentées comme concurrentes, et très utilisées aujourd'hui par les sociologues : la régression logistique et l'analyse des données. Les controverses récurrentes qui les opposent ne sont pas sans rapport avec la distinction suggérée ci-dessus entre, d'une part, les sciences de la nature et de la vie, et, d'autre part, celles de la société et du droit. Pourtant, elles se sont en partie échangées les idées de variables discrètes et continues.

La régression logistique est une extension et une systématisation de l'ancienne idée d'"élimination des effets de structure" ou : "une variable peut en cacher une autre". Cette question a été traitée par la régression multiple et les calculs de corrélation partielle, par Udny Yule, depuis le début du siècle. Mais le problème vient de ce que, dans le cas de la sociologie, les variables dont on souhaite analyser les effets sont souvent "discrètes", c'est-à-dire constituées de "classes d'équivalence", au sens défini ci-dessus. Les modèles de régression logistique (du type LOGIT) permettent d'utiliser des formules de régression linéaire classique par des transformations logarithmiques *ad hoc*. Mais, ce faisant, on revient dans la situation des sciences de la nature où, comme dans les expériences agronomiques de Fisher, on distingue des "effets purs" de variables agissant de façon homogène sur tout l'espace étudié. L'idée que les lois et leurs effets sont transportables et reproductibles, pourvu que soient respectées les conditions *ceteris paribus*, est sous-jacente à cette façon de traiter les variables sociologiques, et elle est issue des sciences de la nature.

Il ne s'agit pas ici de critiquer cet usage, comme cela a déjà été fait maintes fois, depuis les économistes historicistes allemands du XIXe siècle, Simiand, Halbwachs et, plus récemment, Passeron (1991), qui revendique, pour la sociologie, la possibilité d'un "espace non-poppérien du raisonnement", basé sur l'historicité des sociétés humaines. Mais cette nécessaire historicisation n'est pas appliquée par Passeron lui-même aux usages des statistiques à la sociologie. "Historiciser" signifierait étudier, dans un contexte historique donné, la cohérence, formelle et sociale, et l'efficacité propre d'un montage de définitions, de tableaux, de graphiques et de calculs. Ces montages ne peuvent être compris que du point de vue de leur insertion dans un réseau plus vaste d'argumentation et d'action et non pas seulement en tant que porteur d'une connaissance supplémentaire, une brique dans l'édifice de

la science. Un exemple : l'élimination des effets de structure a été pratiquée et discutée au moins depuis les années 1920. Simiand a formulé à leur sujet une critique spectaculaire : "cette méthode conduit à étudier et comparer les comportements d'un renne au Sahara et d'un chameau au Pôle Nord".

Cette boutade a été souvent reprise, jusqu'à nos jours, par ceux qui souhaitent critiquer la transposition du modèle des sciences de la nature aux sociétés humaines. Or cette élimination des effets de structure a été considérablement approfondie et sophistiquée, depuis 1980, par l'usage des modèles de régression logistique, qui permettent précisément de séparer et de quantifier finement les "effets purs" des diverses variables "explicatives". La question n'est donc plus de savoir si ceux qui le font ont raison ou tort, mais *pourquoi* ils le font ? Comment la régression logistique est-elle intégrée dans une plus longue chaîne d'arguments, dans laquelle on peut conjecturer que le *jugement* et *l'action* (et non pas la description) occupent une place centrale. Les débats de épistémologues portent sur ce qu'*il faut faire* pour faire de la "vraie science". Ceux des sociologues des sciences sont différents. Ils portent sur *ce que font* les scientifiques et les objets qu'ils construisent, et pourquoi, sans chercher d'abord à séparer le bon grain de l'ivraie.

Le modèle de la régression logistique est hybride en ce qu'il met en oeuvre des *variables* dites "discrètes", c'est-à-dire découpant exhaustivement l'univers en *classes* disjointes. Les acteurs de son théâtre sont ces variables : ce sont elles qui agissent, ont des effets, purs ou brouillés par ceux de variables concurrentes. Dans les compte-rendus, elles constituent les *sujets des verbes*, et, à ce titre, elles se rattachent au langage des sciences de la nature. Pourtant, au lieu de refléter des mesures, elles rassemblent des classes, constituées sur le modèle des sciences juridiques ou politiques. Mais ces classes ne parlent pas en tant que telles ; elles laissent la parole aux variables : le sexe, l'âge, le diplôme, le revenu, la CSP, la région, la taille de la commune. Ceux qui, à l'image de Karl Pearson et de sa biométrie, sont les plus attirés par le modèle des sciences de la nature, sont gênés par ces variables discontinues. L'âge et le revenu pourraient, à la rigueur, être rapatriés dans le camp des "vraies" variables, mais les autres[3] sont toujours un peu suspectes d'arbitraire et de "conventionnel" : que se passerait-il si on "changeait de nomenclature" ?

Mais le coeur de ces méthodes reste la question des *effets* de certaines variables sur d'autres. Cette interrogation ne trouve sens que dans une perspective d'*action* et de transformation du monde. Sur quoi faut-il agir pour atteindre tel but ? La variable résume alors un objectif (un indicateur social, un critère de convergence fixé par un traité), ou un moyen d'action *de portée générale*. La variable est faite pour être inscrite sur un cadran du tableau de bord de l'homme d'action. La science sociale est

---

3 A l'exception peut être du sexe, mais des essais pour rendre continue sa définition ont été faits par la psycho-sociologie américaine.

une science expérimentale appliquée. Mais elle doit *composer* avec les classes d'équivalence produites historiquement par les Etats de droit : catégories administratives, salariales, scolaires, familiales, fiscales (différentes d'un pays à l'autre, pour le malheur de la construction d'une statistique européenne). C'est pour cette raison que les critiques qui, de Simiand à Passeron, ont visé ces méthodes, ont en partie manqué leur but, et n'ont eu aucun effet. Elles ne s'en prenaient qu'à leur dimension *cognitive*, au lieu de décrire leurs *usages et leurs effets sociaux*, qui ne sont intelligibles que dans une sociologie beaucoup plus vaste des moyens dont dispose une société pour se représenter et agir sur elle-même.

Analyse des données et Data analysis

L'analyse des données dite "à la française", c'est-à-dire issue des travaux de Jean-Paul Benzécri et Brigitte Escoffier, combine, elle aussi, des aspects classificatoires et métrologiques. Elle est d'ailleurs dans la suite directe de l'"analyse factorielle" des psychomètres, qui poursuivaient une démarche typique de la métrologie "symptomatique" des sciences de la vie (Benzécri, 1982). L'intelligence générale (ou "facteur *g*") de Charles Spearman (1904) était une variable latente, "moyenne" des résultats de n épreuves scolaires subies par p élèves. Elle était déterminée comme l'axe principal d'inertie du nuage des p points représentant les performances des élèves dans l'espace à n dimensions des épreuves. L'unidimensionnalité de ce nuage a été ensuite discutée et critiquée par Burt, puis Thurstone, qui cherchent à explorer des axes orthogonaux, décrivant plus fidèlement la complexité de l'espace des "aptitudes". Sans ordinateurs, les psychomètres acquièrent une grande dextérité pour opérer des "rotations d'axes", dans des espaces à beaucoup de dimensions. Surtout utilisée par les psychologues, cette technique est peu connue des sociologues, du moins en France, jusqu'aux années 1960.

Une expérience remarquable resta pourtant isolée et sans suites. En 1954, Jean Porte, le créateur des CSP, effectue à l'Insee une "enquête par sondage sur l'auditoire radiophonique", ancêtre de l'audimat. Dans une "analyse factorielle des goûts "préfigurant, vingt-cinq plus tôt, "*La distinction*" de Bourdieu, il effectue une analyse factorielle d'un tableau des corrélations entre les préférences pour les divers types d'émissions. Il utilise pour cela la méthode de Thurstone dite "centroïde : "Une telle opération ne peut guère être justifiée que par son succès, c'est-à-dire la possibilité d'interpréter les résultats" (Porte, 1954, p. 53). Il interprète le premier facteur comme opposant les "émissions de qualité" aux "émissions légères", puis après une habile rotation d'axes (effectuée graphiquement), un second facteur oppose les "émissions musicales" aux "émissions parlées". Mais l'analyse porte seulement sur les proximités entre émissions, et non sur leurs préférences par les diverses CSP (analysées par des méthodes plus classiques), et surtout l'analyse factorielle ne conduit pas encore à une cartographie, qui sera le propre de l'analyse des correspondances.

Malgré leur homonymie, les méthodes françaises d'*analyse des données*, et les méthodes anglo-saxonnes dites de *data analysis,* popularisées par John Tukey et Eugène Horber, n'ont pas les mêmes philosophies (Deville et Malinvaud, 1983 ; Horber, 1990 ; Schiltz, 1991). Les méthodes anglo-saxonnes distinguent nettement l'analyse *exploratoire,* qui, par des méthodes d'examen et de visualisation très simple d'un fichier, permettent de formuler de premières hypothèses ou des esquisses de modèles probabilistes, testées ensuite par l'*analyse confirmatoire* qui retrouve alors les techniques classiques de la statistique mathématique. En revanche, l'analyse des données à la française se présente comme une fin en soi, en poussant très loin le rejet de tout modèle probabiliste. Elle est avant tout une technique descriptive. Elle ne vise pas à confirmer ou infirmer une théorie préalablement formulée. De ce point de vue, elle renoue avec l'ancienne tradition des sociologues et des économistes historicistes du XIXe siècle, qui bâtissaient des lois "générales" à partir des données observées.

Portant sur des "tableaux de contingence" distribuant des individus selon des classifications variées, l'analyse des correspondances est adaptée à la conception "conventionnelle", issue des sciences politiques et du droit. Elle distribue ces classes selon des systèmes de proximités, possédant des configurations de propriétés voisines. Dans ce cas, les acteurs du théâtre ainsi mis en scène sont des *groupes* (ou même des *individus*), et non plus des *variables*. Les sujets des verbes sont, dans les phrases des interprétations, ces groupes (qui peuvent être définis par le sexe, l'âge, la CSP , etc.). Ceux-ci ont une existence autonome par rapport à la nomenclature exhaustive (à la différence des méthodes de régression logistique). Ces méthodes peuvent servir de façon classificatoire *a posteriori,* en regroupant (de façon ascendante) des individus, ou en découpant (de façon descendante) l'ensemble initial, après définition d'une "distance", minimisée à l'intérieur des classes et maximisée entre les classes. Dans ce cas, l'analyse statistique engendre littéralement de nouvelles formules *d'équivalences conventionnelles*, réutilisables pour l'action, et n'ayant d'autre portée que dans l'usage qui en est fait.

Mais, dans sa version "cartographique", très utilisée, l'analyse des correspondances retrouve la perspective métrologique et les variables latentes. Les "axes d'inertie", déterminés par diagonalisation des matrices de variance covariance, engendrent un nouvel espace, dans lequel les individus et les groupes ont des "coordonnées". Il est tentant d'interpréter celles-ci, c'est-à-dire de les traiter comme des mesures continues de "quelque chose" qui, bien que non directement visible, existerait dans la nature. Certaines interprétations de Benzecri, associant parfois la structure des axes à un dessein divin, rappellent irrésistiblement celles de Quetelet, pour qui l'"homme moyen" ne pouvait être que le produit de la volonté divine. Qu'il s'agisse simplement de la nature, ou de Dieu, une statistique réaliste peut toujours contribuer à engendrer du réel, par la seule efficacité des ses procédures de calcul et d'objectivation.

Ainsi, chacune à leur façon, la régression logistique et l'analyse des correspondances opèrent une hybridation entre les optiques métrologiques et classificatoires. Elles constituent aujourd'hui deux des méthodologies statistiques les plus utilisées par les sociologues. On ne peut cependant les comparer, tant leurs langages et leurs usages sont différents[4] . Surtout, elles sont utilisées par des sociologues et dans des contextes institutionnels très distincts, ce qui rend difficile une confrontation *sociologique* de leurs différences d'usages. Les produits des régressions logistiques sont présentés comme des *résultats*, associant des effets à des causes, portant sur des variables décontextualisées et supposées de portée générale, à la façon dont les sciences expérimentales déroulent les étapes de leurs investigations (Licoppe, 1996). De ce point de vue, ils semblent au coeur de la démarche scientifique d'une sociologie qui progresse en accumulant de tels résultats.

En revanche, l'analyse des données à la française est rarement présentée (à la différence de la *data analysis* anglaise) comme préalable à une "analyse confirmatoire", vérifiant des hypothèses théoriques dont elle serait une des sources. Elle est plutôt un élément parmi d'autres d'un ensemble de descriptions historiques de la complexité et des dimensions d'un univers social. Les "variables" ne figurent pas en tant que telles, mais à travers les classes qu'elles distinguent. Ce sont les configurations singulières de ces classes et de leurs propriétés qui font l'objet du commentaire du sociologue. La généralisation éventuelle procède d'une rhétorique différente de celle des sciences de la nature ou de la vie. C'est la juxtaposition de configurations similaires qui fournit un argument. Ainsi, la structure bi-dimensionnelle de l'espace des catégories sociales françaises a été suggérée et confirmée par une succession de travaux analysant les comportements de ces catégories à divers points de vue : structure des consommations, pratiques culturelles, distribution spatiale dans les quartiers urbains, intermariages, comportements électoraux (Desrosières, Thévenot, 1988). Ces configurations sont historiques en ce qu'elles dépendent de taxinomies, plus ou moins durcies et elles-même historiques, et de pratiques dont le sens évolue.

Ces différences d'usage reflètent le relatif émiettement d'une discipline, la sociologie, qui tire sa légitimité (mais peut être aussi son originalité), d'un patchwork de modèles de scientificité. Elle aurait peut être à gagner à *expliciter* ce mélange et sa portée sociologique, en termes d'insertion de son discours dans des pratiques sociales différentes, plutôt qu'à chercher à faire triompher l'un ou l'autre de ces modèles. L'histoire montre que ces combats en apparence "épistémologiques", sont en général sans issue, car chacun de ces modèles a un usage social déterminé. Les remarques qui précèdent ne sont d'ailleurs que des hypothèses, qui demanderaient à être validées par une recherche détaillée et comparative sur les usages sociaux des méthodologies statistiques en sociologie, selon les institutions et selon les pays.

---

4 Une discussion de ces différences est proposée par Félicité Hay des Nétumières (1996).

BIBLIOGRAPHIE

ARMATTE M., 1995 : *Histoire du modèle linéaire. Formes et usages en économie et économétrie jusqu'en 1945.* Thèse de doctorat, EHESS, Paris.

BENZECRI J.P., 1982 : *Histoire et préhistoire de l'analyse des données,* Dunod, Paris.

BOLTANSKI L., 1970 : "Taxinomies populaires, taxinomies savantes : les objets de consommation et leur classement", *Revue française de sociologie,* XI, pp. 34-44.

BOLTANSKI L., THEVENOT L. 1983 : "Finding One's Way in Social Space ; a Study Based on Games", *Social Science Information,* vol. 22, 4-5, pp. 631-679.

CICOUREL A., 1964 : *Method and Measurement in Sociology,* The Free Press of Glencoe, New York.

CLOGG C., 1992 : "The impact of sociological methodology on statistical methodology", *Statistical Science,* vol. 7, n° 2, pp. 183-207.

DESROSIERES A., 1976 : "Eléments pour l'histoire des nomenclatures socioprofessionnelles", in *Pour une histoire de la statistique, tome 1 : contributions,* réédité en 1987, Economica-Insee, pp. 155-231.

DESROSIERES A., 1993 : *La politique des grands nombres. Histoire de la raison statistique,* La Découverte, Paris.

DESROSIERES A., THEVENOT L., 1988 : *Les catégories socioprofessionnelles,* La Découverte, Paris.

DEVILLE J.C., MALINVAUD E, 1983 : "Data Analysis in Official Socio-economic Statistics", *Journal of the Royal statistical Society,* A, vol. 146, Part 4, pp. 335-361.

FAGOT-LARGEAULT A., 1989 *: Les causes de la mort. Histoire naturelle et facteur de risque,* Vrin, Paris.

GOULD S.J., 1983 : *La Mal-mesure de l'homme,* Ramsay, Paris.

GREENACRE M., BLASIUS J. (eds), 1994 *: Correspondance Analysis in the Social Sciences,* Academic Press, New York, London.

GUIBERT B., LAGANIER J., VOLLE M., 1971 : "Essai sur les nomenclatures industrielles", *Economie et Statistique*, 20, pp. 23-36.

HAAVELMO T., 1944 : "The Probability Approach in Econometrics", *Econometrica,* 12, pp. 1-118.

HAY des NETUMIERES F., 1996 : "Méthodes de régression et analyse factorielle", in *Construction de l'objet et modèlisation en sciences sociales,* Université René Descartes, Paris V-Sorbonne.

HERAN F., 1984 : "L'assise statistique de la sociologie", *Economie et Statistique,* 168, pp. 23-35.

HORBER E., 1990 : *Analyse exploratoire des données et sciences sociales. Vers une approche méthodologique pragmatique.* Thèse de doctorat, Université de Genève.

KRAMARZ F., 1991 : "Déclarer sa profession", *Revue française de Sociologie,* XXXII, pp. 3-27.

LICOPPE C., 1996 : *La formation de la pratique scientifique. Le discours de l'expérience en France et en Angleterre (1630-1820),* La Découverte, Paris.

MAC KENZIE D., 1981 : *Statistics in Britain, 1865-1930. The Social Construction of Scientific Knowledge,* Edinburgh University Press, Edimburgh.

MERLLIE D., 1990 : "Les classements professionnels dans les enquêtes de mobilité", *Annales* ESC, XLV, n° 6, nov-déc.

MORGAN M., 1990 : *The History of Econometric Ideas,* Cambridge University Press, Cambridge.

PASSERON J.C., 1991 : *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel,* Nathan, Paris.

PORTE J., 1954 : "Une enquête par sondage sur l'auditoire radiophonique", *Bulletin mensuel de Statistique,* supplément janvier-mars 1954, Insee, pp. 31-58.

ROBERT P. et alii, 1994 : *Les comptes du crime. Les délinquances en France et leurs mesures.* L'Harmattan-Paris.

SALAIS R., BAVEREZ N., REYNAUD B., 1986 : *L'invention du chômage,* PUF, Paris.

SCHILTZ M.A., 1991 : "Influence du choix des traitements statistiques sur les opérations élémentaires dans un dépouillement d'enquête : hypothèses, codage et sélection des variables". Communication au Congrès de l'Association internationale de sociologie, juillet 1990, Madrid.

SPEARMAN C., 1904 : "General Intelligence Objectively Determined and Measured", *American Journal of Psychology,* 15, pp. 201-293.

STIGLER S., 1986 : *The History of Statistics. The Measurement of Uncertainty before 1900,* Harvard University Press, Cambridge (Mass).

Conférences spéciales

CONFIDENTIALITÉ DES DONNÉES OU L'ART DU BROUILLAGE CLAIR

Jean-René Boudreau[1]

1. Introduction

Nous sommes bel et bien à l'ère de l'information. Les sondages sont présentement en vogue. Les sociétés deviennent de plus en plus complexes. Elles ont le besoin, pour mesurer leur pouls, d'avoir une information de haut niveau. Par exemple, Statistique Canada doit constamment réajuster le tir afin de fournir les produits de diffusion les plus pertinents pour les décideurs. Ces réajustements se font au moins sur trois plans. Premièrement, les lois et politiques canadiennes (qu'il suffise de mentionner l'immigration, le multiculturalisme, l'équité en matière d'emploi) requièrent des données ciblant des sous-groupes de la population canadienne. Les utilisateurs veulent également que l'information soit localisée, c'est-à-dire présentée à des niveaux géographiques très riches. Deuxièmement, les supports de l'information sont ajustés pour permettre un traitement efficace des données. Historiquement, le support papier était roi et maître. Puis, vint l'avènement de la bande magnétique et plus récemment celui du disque laser (CD-ROM). Ces supports offrent les avantages d'une grande compacité et d'une recherche plus efficace d'information. Troisièmement, l'accès aux données est amélioré en utilisant les réseaux d'information tel Internet et en développant des logiciels conviviaux pour permettre aux utilisateurs de soumettre eux-mêmes leurs requêtes. Toutes ces mesures ont pour but une plus grande disponibilité de la masse d'information que les enquêtes recueillent auprès des répondants. Cette disponibilité se définit en termes de rapidité de diffusion autant qu'en termes du nombre d'utilisateurs. Cette dynamique appelle un consensus entre diffuser une information de plus en plus pertinente et obtempérer aux articles de la Loi de la Statistique du Canada qui obligent de maintenir confidentielles les réponses des canadiens et canadiennes.

Le risque de divulgation d'un produit de diffusion est une évaluation de la possibilité, pour un individu ou pour une entreprise, d'établir la provenance d'une information recueillie par l'agence statistique. Nous sommes d'avis qu'à toute diffusion d'information, il y a un risque de divulgation apparenté. L'agence se trouve donc confrontée à deux problèmes : trouver une bonne façon d'estimer le

1. M. Boudreau est méthodologiste principal à la Division des méthodes d'enquêtes sociales de Statistique Canada, Ottawa, Canada, K1A 0T6. Les opinions exprimées dans cet article sont celles de l'auteur et ne reflètent pas nécessairement celles de Statistique Canada.

risque de divulgation et déterminer des méthodes de réduction de ce risque lorsque ce dernier est jugé trop élevé. Les règles de confidentialité sont des actions prises sur les données qui permettent une diffusion avec un risque de divulgation moins élevé. Quelles sont ces règles ? Seulement deux types d'actions sont opérées sur les données pour garantir un haut niveau de confidentialité : la suppression de données et l'introduction de bruit dans les données. La suppression de données est facile à expliquer : on refuse tout simplement de diffuser une partie de l'information. L'introduction de bruit est un ensemble de méthodes permettant de modifier les données sans enlever leur caractère statistique. La suppression de données intervient à différents niveaux. On peut choisir de supprimer un tableau entier ou une variable dans un fichier de microdonnées (suppression globale) ; ou, on peut choisir de supprimer une ou plusieurs cellules d'un tableau ou une ou plusieurs catégories d'une variable pour certains enregistrements dans un fichier de microdonnées (suppression locale). L'arrondissement des valeurs est aussi une méthode d'introduction de bruit dans les données pour les variables de quantité d'un fichier de microdonnées, de même que le regroupement de catégories d'une variable ordinale ou codifiée. Les règles de confidentialité doivent se faire le plus discrètement possible. D'une part, les utilisateurs et le public en général doivent voir qu'il y a eu un certain traitement dans les données. Mais d'autre part, la substance statistique des données doit rester intacte. Il faut absolument que les analyses faites à partir des données puissent toujours être valides. D'où le titre de l'article. On doit créer du brouillage qui doit être visible mais sans gêner la vision.

Nous nous concentrerons sur les produits de diffusion appelés "fichiers de microdonnées". Nous commençons par décrire une façon d'évaluer le risque de divulgation d'un fichier de microdonnées. Cette évaluation dépend du calcul d'une probabilité conditionnelle bien particulière. Ainsi la deuxième section traite de la formalisation et d'une modélisation possible de cette probabilité. L'auteur, en outre, suggère un modèle empirique qui colle bien avec la réalité. Après avoir étudié et évalué le risque, nous proposons dans la troisième section un traitement de données pour réduire le risque de divulgation. Nous discuterons des méthodes d'introduction de bruit les plus utilisées : soit l'échange de données (data swapping) et la suppression de valeurs. Nous établissons une formule inédite du biais créé par un échange de données. Nous donnons également, avec l'aide d'un nouveau concept appelé "multiplicité d'un enregistrement", une façon de déterminer les enregistrements les plus dangereux. Pour terminer, nous donnons une méthode d'introduction de bruit lorsqu'il y a des variables de quantité comme des sources de revenu. Mais avant de procéder au traitement, nous devons premièrement évaluer le risque de divulgation d'un fichier de microdonnées.

2. Évaluation du risque de divulgation

2.1. Discussion de la problématique

Considérons le problème d'appariement suivant. Un échantillon aléatoire simple est tiré d'une population (fichier A). Nous voulons apparier ce fichier avec un autre fichier (fichier B) provenant de cette même population en utilisant toutes les variables ordinales ou codifiées communes aux deux fichiers. Ces variables seront appelées discrètes[2] par la suite. Le cas où certaines variables en commun ne sont pas discrètes sera discuté plus tard. Nous supposons que les erreurs de saisie et de réponse sont négligeables. Si un appariement biunivoque est obtenu entre deux enregistrements, quel "niveau de confiance" peut-on accorder à l'énoncé : « ces deux enregistrements proviennent de la même unité de la population » ? Ce niveau de confiance nous aide à évaluer le risque de divulgation du fichier A. En effet, si une agence statistique diffuse un fichier de microdonnées (fichier A), certains individus ou organismes pourraient tenter de coupler leurs propres fichiers de microdonnées (ex : fichier B) à celui de l'agence dans le but d'identifier la provenance de certains enregistrements (en utilisant les variables discrètes communes aux deux fichiers). Ainsi, l'agence statistique doit s'assurer que le niveau de confiance sera le plus bas possible avant la diffusion du fichier, ceci afin d'éliminer toute incitation à coupler le fichier diffusé avec d'autres fichiers.

Une condition nécessaire pour avoir un haut niveau de confiance est d'imposer que le fichier B couvre bien la population ou la sous-population d'intérêt. Sous cette hypothèse, le niveau de confiance — que nous assimilons maintenant au risque de divulgation — est intimement relié à la probabilité conditionnelle d'être un élément unique dans la population (par rapport aux variables d'appariement) étant donné d'être un élément unique dans l'échantillon. Mais cette probabilité conditionnelle n'est pas le risque de divulgation. Deux autres facteurs sont à considérer : la détérioration des variables et la possibilité que de tels fichiers B existent. Le premier facteur réduit le pouvoir d'identification des fichiers. En effet, les problèmes de couverture, d'erreur de réponse, de non-réponse, d'actualisation des valeurs des variables, etc... ne peuvent que réduire la confiance que nous pourrions avoir face à la véracité d'un couplage entre deux enregistrements. L'autre facteur est encore plus important. Le risque de divulgation est la possibilité d'établir un lien et d'y croire. En gros, cette possibilité est une somme pondérée de probabilités conditionnelles. Nous nous expliquons. La possibilité [probabilité] d'établir un lien peut s'écrire comme :

$$Risque = \int P(\textit{unique population} \mid \textit{unique échantillon} ; \textit{contenu fichier B})\, P(\textit{Contenu fichier B})$$

2. Les variables qui ne sont pas discrètes sont appelées "réelles" (c'est-à-dire qu'elles représentent une quantité, une magnitude). Une source de revenu en est une, par exemple.

Cette équation s'interprète comme suit. Pour trouver le risque de divulgation d'un fichier A, il faut faire la somme des probabilités conditionnelles (qui dépendent du contenu des fichiers A et B : les variables en commun aux deux fichiers) multipliées par les possibilités [probabilités] qu'il existe à l'extérieur de l'agence de tels fichiers B. En clair, cela veut dire que si vous prenez beaucoup de variables d'appariement, la probabilité conditionnelle sera très élevée mais la possibilité d'avoir un tel fichier sera sans doute négligeable ou même inexistante ; le risque de divulgation [la somme pondérée] en sera peut-être également négligeable. Cette pondération ne peut pas être estimée statistiquement mais elle peut et doit être évaluée par des personnes connaissant l'ensemble des fichiers externes à l'agence. Tout ce que l'on peut faire est de déterminer les probabilités conditionnelles pour différents contenus et de se souvenir de toujours pondérer les résultats.

L'estimation du nombre d'éléments uniques dans la population a fait l'objet de beaucoup de recherches ces dernières années. Greenberg et Zayatz[3] donnent deux façons d'estimer le nombre d'éléments uniques. La première consiste à ré-échantillonner l'échantillon selon le même plan de sondage. L'estimateur est construit en supposant que les relations entre les éléments uniques de la population et du premier échantillon sont les mêmes entre celles du premier et du deuxième échantillon. La deuxième façon proposée par ces auteurs utilise la structure de la population, c'est-à-dire la description de la population en termes du nombre de cellules définies par les variables d'appariement ayant exactement une unité, deux unités, etc... Ce qu'ils appellent « classes d'équivalence ». Ces deux techniques donnent de bons résultats si la fraction de sondage est supérieure à 10 %. Une autre façon de procéder est d'essayer de modéliser la structure de la population et d'estimer les paramètres à partir d'un échantillon. Bethlehem, et cie.[4] ont tenté de modéliser la proportion du nombre d'éléments uniques dans la population à l'aide d'un modèle dérivé de la loi Poisson-Gamma. Ce modèle souffre d'un manque d'ajustement important. Skinner et Holmes[5] ont modélisé la proportion d'uniques dans la population en utilisant la loi Poisson-Lognormal. Ils obtiennent des résultats qui collent beaucoup plus à la réalité. L'auteur propose d'utiliser la théorie de l'échantillonnage pour déterminer exactement la forme de la relation entre les éléments uniques dans la population et ceux dans l'échantillon lorsque les variables d'appariement sont toutes discrètes. Nous essayerons de modéliser cette relation pour de petites fractions de sondage. Nous donnerons par la suite un exemple d'évaluation du risque de divulgation.

3. Greenberg B. V., Zayatz (1992). Strategies for Measuring Risk in Public Use Microdata Files. Statistica Neerlandica.

4. Bethlehem, J. G., Keller, W. J., Pannekeok, J., (1990). Disclosure Control of Microdata. JASA, 85, pp. 38-45.

5. Skinner C. J.,Holmes D. J. (1992). Modelling Population Uniqueness. International Seminar on Statistical Confidentiality, Dublin.

2.2 Détermination de la probabilité conditionnelle

Nous avons une population de N éléments ou unités. Le contenu, c'est-à-dire les variables d'appariement, partitionne cette population en m sous-populations de taille $N_1, \ldots, N_m$. La structure de la population est donnée par le vecteur $(U_1, \ldots, U_N)$ où $U_j = \text{card}\ \{\ k : N_k = j\ \}$. Nous prenons un échantillon de taille n tiré d'une manière aléatoire simple de cette population. Nous observons le vecteur aléatoire $(n_1, \ldots, n_m)$, dont les composantes sont respectivement le nombre d'unités échantillonnées de la sous-population k (k = 1, ... ,m). La structure de l'échantillon est le vecteur aléatoire $(u_1, \ldots, u_n)$ où $u_j = \text{card}\ \{\ k : n_k = j\ \}$. Un élément sera dit unique dans la population s'il appartient à une sous-population de taille unité. Une unité échantillonnée sera dite unique dans l'échantillon si elle est la seule unité échantillonnée à appartenir à sa sous-population. Puisqu'un élément unique dans la population qui est échantillonné est nécessairement unique dans l'échantillon, nous obtenons que la probabilité conditionnelle d'être unique dans la population étant donné d'être unique dans l'échantillon est le rapport entre les proportions des éléments uniques dans la population et ceux dans l'échantillon. Donc, nous voulons avoir une estimation de

$$P = f\frac{U_1}{E\{u_1\}}$$

où f est la fraction de sondage et l'espérance mathématique est celle établie par le plan de sondage. L'espérance est nécessaire pour obtenir un paramètre au niveau de la population. Ce paramètre, par abus de langage, sera tout de même considéré comme une probabilité conditionnelle. Elle n'est pas loin de l'idée du risque de divulgation ou du niveau de confiance expliqué précédemment. Nous avons un premier résultat.

Théorème A. *Si un échantillon aléatoire simple de taille n est tiré d'une population de taille N possédant la structure $(U_1, \ldots, U_N)$, alors*

$$E\{u_j\} = \frac{\binom{N-j}{n-j}}{\binom{N}{n}} U_j + \sum_{i=1}^{\infty} \frac{\binom{j+i}{j}\binom{N-j-i}{n-j}}{\binom{N}{n}} U_{j+i}\ .$$

Démonstration. La somme est en réalité finie. Puisque u_j est à valeurs entières, nous pouvons utiliser l'identité

$$E\{u_j\} = \sum_{i=1}^{\infty} P\{u_j \geq i\} .$$

Posons $A_k = \{ (n_1, \ldots, n_n) : n_k = j \}$. Nous avons l'identité suivante

$$P\{u_j \geq i\} = P\left\{ \underset{k_1 < \ldots < k_i}{U} A_{k_1} \ldots A_{k_i} \right\}$$

Nous pouvons montrer facilement que

$$\sum_{i=1}^{\infty} P\{u_j \geq i\} = \sum_{k=1}^{m} P\{A_k\} .$$

En effet, il suffit de déterminer la probabilité de chaque union et de réaliser que tous les termes s'annulent sauf la somme des probabilités des événements A_k. Maintenant, $P\{ A_k \}$ vaut

$$P\{A_k\} = \frac{\binom{N_k}{j}\binom{N - N_k}{n - j}}{\binom{N}{n}} .$$

Donc l'espérance de u_j vaut

$$E\{u_j\} = \sum_{\substack{k=1 \\ j \leq N_k \leq N - n + j}}^{m} \frac{\binom{N_k}{j}\binom{N - N_k}{n - j}}{\binom{N}{n}} = \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N - i}{n - j}}{\binom{N}{n}} U_i .$$

Ce qu'il fallait démontrer.

En particulier pour j = 1, nous avons

$$E\{u_1\} = fU_1 + \sum_{i=1}^{\infty} (i + 1) \frac{\binom{N - 1 - i}{n - 1}}{\binom{N}{n}} U_{1+i} ,$$

qui peut s'écrire comme :

$$E\{u_1\} = f\,U_1 + n \sum_{i=1}^{N-n} \frac{(i+1)}{N-i} \left(1 - \frac{n}{N}\right) \ldots \left(1 - \frac{n}{N-i+1}\right) U_{1+i}$$

$$\approx f\left(U_1 + \sum_{i=1}^{N(1-f)} (i+1)(1-f)^i\, U_{1+i}\right)$$

si N est suffisamment grand. Ainsi la probabilité conditionnelle que nous recherchons devient

$$P = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)(1-f)^i \frac{U_{1+i}}{U_1}} .$$

Cette fonction donne pour f = 0 la proportion d'uniques dans la population. Un examen du développement de Taylor autour de l'origine nous renseigne sur la concavité de la relation dans le domaine des petites fractions de sondage pour des structures de population réelles.

2.3. *Modélisation de la relation entre P et f*

Au cours des dernières années, plusieurs personnes ont tenté avec plus ou moins de succès de modéliser la structure de la population (en particulier le nombre d'éléments uniques dans la population). La première tentative connue de l'auteur est celle de Bethlehem et cie[6]. Ils ont supposé, sans justification outre celle de la simplicité des techniques, que la structure d'une population pourrait être simulée à partir d'un modèle Poisson-Gamma. Sous cette hypothèse, on en vient facilement à trouver une expression paramétrique pour la proportion d'éléments uniques dans la population. En fait, l'expression est donnée par

$$E_m\{U_1/N\} = \left(\frac{1}{1+\beta N}\right)^{1+\alpha}$$

6. op.cit.

où α et β sont les paramètres de la loi Gamma du modèle (α, $\beta > 0$). Dès que l'on essaie, à partir d'un échantillon, d'estimer ces paramètres, on s'aperçoit vite que le modèle souffre d'un manque d'ajustement. Le paramètre α est invariablement estimé à une valeur non significativement différente de zéro. Même les techniques classiques de compensation ne permettent pas de stabiliser le modèle. Nous verrons plus loin que le problème se situe au niveau de l'étendue de l'intervalle de définition de α. Pour sa part, Skinner et Holmes[7] proposent une approche basée sur la théorie de la classification. Selon cette théorie, la structure de la population serait simulée par un modèle Poisson-Lognormal. Ce modèle est beaucoup plus difficile à maîtriser que celui énoncé précédemment. Par contre, les résultats obtenus semblent très bien coller à la réalité. L'approche que nous développons dans cet article modélise directement la relation entre P et f au lieu de modéliser la structure sous-jacente. Cette approche, de nature purement empirique, a l'avantage de coller à la réalité si on est en mesure de pouvoir observer un grand nombre de populations réelles. Un désavantage toutefois de cette méthode est qu'elle ne donne aucun renseignement sur comment ces populations sont générées. Autrement dit, elle ne donne aucune justification théorique ni n'en suggère.

Cette approche empirique ne suppose aucune hypothèse probabiliste sauf celle de la sélection d'un échantillon aléatoire simple. La méthode consiste à étudier la relation entre P et f pour plusieurs populations obtenues par voie de recensements, de tenter de décanter les ressemblances, de proposer une formulation paramétrique de P en fonction de f, et de proposer une méthode d'estimation des paramètres en utilisant un échantillon. La formulation de cette relation qui colle très bien avec l'observation est donnée par l'expression suivante

$$P_M = E_m \{ P \} = \left(\frac{f + \gamma}{1 + \gamma} \right)^{\alpha},$$

où $0 < \alpha < 1$ et $\gamma > 0$. Le paramètre α influe directement sur la concavité observée de la relation. Nous notons que le modèle Poisson-Gamma implique une relation convexe entre P et f. Mais nous savons par observation que la relation est concave. Ainsi, vouloir ajuster le modèle Poisson-Gamma à la réalité ne peut que donner quelque chose proche de la linéarité ($\alpha = 0$). C'est exactement ce que l'on lit dans la littérature. En pratique, nous ne pouvons pas utiliser directement la relation entre P et f pour estimer les paramètres α et γ puisque la probabilité conditionnelle n'est pas observable. Les seules quantités d'intérêt observables sont les composantes de la structure de l'échantillon. Essayons de dériver l'espérance du nombre d'éléments uniques dans un échantillon à partir de P et f.

7. op. cit.

Théorème B. *Si la formulation paramétrique entre P et f est correcte avec paramètres α et γ, alors l'espérance, au sens du modèle, de la proportion du nombre d'éléments uniques dans l'échantillon est donnée par*

$$Q_n = E_m \{ u_1 / n \} = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha},$$

β est le réciproque de la multiplication de γpar la taille de la population.

Démonstration : Par définition, la probabilité conditionnelle recherchée est le quotient des proportions d'éléments uniques dans la population et dans l'échantillon respectivement. Puisque la formulation entre P et f est correcte, nous avons

$$P_M = \left(\frac{\frac{n}{N} + \gamma}{1 + \gamma} \right)^{\alpha} = \left(\frac{1 + \frac{n}{\gamma N}}{1 + \frac{N}{\gamma N}} \right)^{\alpha} = \left(\frac{1 + \beta n}{1 + \beta N} \right)^{\alpha} = \frac{Q_N}{Q_n}.$$

Ce qui donne

$$Q_n = K \left(\frac{1}{1 + \beta n} \right)^{\alpha}.$$

Puisque $Q_1 = 1$, nous obtenons le résultat recherché.

Nous pouvons nous demander si la relation entre Q_n et n colle à la réalité. Nous avons pris comme exemple une population réelle de taille 800 000 avec un contenu de quatre variables de laquelle nous avons sélectionné un premier échantillon avec une fraction de sondage de 0,005. À partir de cet échantillon, nous avons sélectionné d'une manière complètement indépendante 900 échantillons : 100 échantillons avec une fraction de sondage de 0,1 ; 100 avec une fraction de 0,2 ; ... ; 100 échantillons avec une fraction de 0,9. Puisque l'échantillon premier et les autres sont tirés selon le plan aléatoire simple, tous ces échantillons sont tirés par un plan aléatoire simple (seule la fraction de sondage change). Le *tableau suivant* donne les différentes valeurs de Q_n en plus des variations observées des proportions. Ces résultats collent très bien avec la théorie.

Tableau 1
Moyenne d'éléments uniques dans l'échantillon selon la taille d'échantillon

Taille d'échantillon (n)	Moyenne d'éléments uniques dans l'échantillon (Q_n)	Écart-type de la moyenne des éléments uniques dans l'échantillon	Taille d'échantillon (n)	Moyenne d'éléments uniques dans l'échantillon (Q_n)	Écart-type de la moyenne des éléments uniques dans l'échantillon
390	0,503	0,0261	2 345	0,274	0,0073
781	0,408	0,0200	2 736	0,258	0,0062
1 172	0,356	0,1577	3 127	0,244	0,0049
1 563	0,321	0,0113	3 518	0,233	0,0038
1 954	0,294	0,0092			

Il ne reste qu'à trouver une méthode d'estimation des paramètres. Le modèle s'écrit de la manière suivante

$$u_1/n = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha} + \varepsilon,$$

où l'erreur ε est régie par la loi des écarts des u_1. Les méthodes standards d'estimation des paramètres de la sorte dépendent lourdement de cette loi. Puisque nous ne la connaissons pas et nous ne sommes pas en mesure d'émettre des hypothèses, nous allons plutôt utiliser les réalisations des moyennes Q_n et de supposer que ces points seront près de la courbe si les moyennes sont basées sur plusieurs expériences (par ex : 100 échantillons). La méthode est la suivante :

I. Sélectionner des échantillons aléatoires simples répétés de l'échantillon original selon plusieurs fractions de sondage (ex : 0,1, 0,2, ... , 0,9). Le nombre de répétitions pour chacune de ces fractions de sondage doit être élevé.

II. Pour chacune des fractions de sondage, calculer les moyennes du nombre d'éléments uniques dans l'échantillon (Q_n).

III. Utiliser une méthode numérique[8] pour déterminer les paramètres α et β qui collent le plus à l'observation.

8 Nous avons utilisé l'algorithme NEWTON programmé dans la procédure NLIN du progiciel SAS (version 6.10).

IV. Déterminer γ à partir de β.

V. Calculer les probabilités conditionnelles à partir du modèle.

2.4. Exemple d'évaluation du risque

Dans cette sous-section, nous allons calculer des probabilités conditionnelles du fichier de microdonnées à grande diffusion du recensement de la population canadienne de 1991. Dans une première étape, nous calculons la vraie probabilité conditionnelle (si nous ignorons la détérioration des variables) à partir de la formulation trouvée. Nous prenons comme contenu l'ensemble des variables avec des données recensées disponibles sur le fichier de microdonnées à grande diffusion qui nous paraissent discriminantes. Elles sont au nombre de neuf : la province(11)[9], la région métropolitaine du recensement(20), le nombre de personnes dans le ménage(8), la langue maternelle(18), l'âge simple, le sexe(2), l'état matrimonial(5), le statut de la famille du recensement(13) et le mode d'occupation du logement(2). Même si cette première analyse se restreint à un nombre limité de variables, celles-ci sont cependant les plus populaires, c'est-à-dire qu'elles se retrouvent le plus souvent sur d'autres fichiers de microdonnées. Elles sont donc plus susceptibles d'être utilisées comme variables d'appariement. La possibilité d'avoir un tel fichier ne peut pas être considérée comme négligeable même si elle est faible. Ainsi, les résultats obtenus seront utiles pour évaluer le risque que nous prenons à la diffusion des fichiers de microdonnées du recensement.

La population à l'étude est l'ensemble de la population canadienne. Le tableau croisé de ces neuf variables pour toute la population définit le contenu. Nous utilisons la formule approximative de la probabilité conditionnelle pour déterminer les valeurs de P pour différentes fractions de sondage. Les paramètres de la modélisation de la relation entre la probabilité conditionnelle et la fraction de sondage, obtenus par la méthode des moindres carrés, sont 0,598251 et 0,006476 respectivement pour α et γ. Les résultats sont donnés au *tableau suivant* :

9 La notation "variable (k)" indique que la variable possède k valeurs.

Tableau 2
Probabilité conditionnelle et modélisation pour le contenu de neuf variables

Fraction de sondage (f)	Probabilité conditionnelle d'après la formule (P)	Probabilité conditionnelle d'après le modèle (P_M)	Fraction de sondage (f)	Probabilité conditionnelle d'après la formule (P)	Probabilité conditionnelle d'après le modèle (P_M)
0,0001	0,032	0,049	0,0950	0,263	0,253
0,0005	0,039	0,051	0,1000	0,270	0,261
0,0010	0,045	0,053	0,1500	0,334	0,328
0,0050	0,074	0,069	0,2000	0,390	0,388
0,0100	0,096	0,085	0,2500	0,441	0,441
0,0150	0,113	0,100	0,3000	0,488	0,491
0,0200	0,127	0,113	0,3500	0,533	0,537
0,0250	0,140	0,126	0,4000	0,576	0,581
0,0300	0,152	0,137	0,4500	0,617	0,623
0,0350	0,163	0,148	0,5000	0,656	0,663
0,0400	0,173	0,159	0,5500	0,694	0,702
0,0450	0,182	0,169	0,6000	0,731	0,739
0,0500	0,192	0,178	0,6500	0,768	0,774
0,0550	0,201	0,188	0,7000	0,803	0,809
0,0600	0,209	0,197	0,7500	0,837	0,843
0,0650	0,217	0,206	0,8000	0,871	0,876
0,0700	0,226	0,214	0,8500	0,904	0,908
0,0750	0,233	0,222	0,9000	0,937	0,939
0,0800	0,241	0,230	0,9500	0,969	0,970
0,0850	0,248	0,238	1,0000	1,000	1,000
0,0900	0,256	0,246			

La proportion d'éléments uniques dans la population s'établit à 2,9 %. Si on considère la fraction de sondage utilisée pour les fichiers de microdonnées de 1991, c'est-à-dire 3 %, la probabilité conditionnelle obtenue pour ce contenu de neuf

variables est de 15 %. Ce résultat doit être pondéré par la possibilité d'avoir un fichier pouvant contenir ces variables.

Comme, pour cette analyse, nous possédions ces neuf variables pour toute la population, nous étions en mesure d'évaluer la probabilité conditionnelle réelle pour ce contenu. Nous ajoutons à ce contenu certaines variables discriminantes obtenues par échantillonnage qui se trouvent dans le fichier de microdonnées. Ces variables sont : l'origine ethnique(33), le plus haut certificat ou diplôme(14), la profession(14) (selon la classification de 1991), l'industrie(16) (selon la classification type des industries de 1980) et le revenu total(11). Nous avons seulement comme donnée la structure de l'échantillon du recensement. La fraction de sondage est 20 %. Cependant, avec cette fraction élevée, il est possible de remplacer dans la formule qui suit le théorème A la structure de la population par celle de l'échantillon. En effet, pour des fractions supérieures à 10 %, la statistique obtenue p est très proche de la probabilité conditionnelle. Cela n'est pas vrai pour les fractions de sondage plus petites que 10 %. En fait p converge vers 1 lorsque f tend vers 0. La valeur de p pour une fraction de 20 % est 0,83. En utilisant cette estimation de la probabilité conditionnelle et en connaissant la proportion d'uniques dans l'échantillon du recensement, nous trouvons que la proportion d'uniques dans la population se situe à 0,51. Si nous divisons ce nombre par la proportion d'uniques dans l'échantillon du fichier de microdonnées, nous trouvons une estimation de la probabilité conditionnelle avec le contenu augmenté du fichier de microdonnées : P = 0,66. Si notre estimation est bonne, le fait d'ajouter ces 5 nouvelles variables au contenu de neuf variables augmenterait de façon très importante la probabilité conditionnelle. Pour vérifier si cette estimation est raisonnable, nous avons estimé cette probabilité en utilisant une autre méthode.

Nous allons utiliser la modélisation de P. Comme nous ne connaissons pas la structure de la population pour ce contenu (14 variables), nous ne pouvons pas utiliser la relation entre P et f pour estimer les paramètres α et β du modèle. Les seules quantités d'intérêt observables sont les composantes de la structure de l'échantillon. Si la formulation paramétrique entre P et f est correcte, alors la proportion d'éléments uniques dans l'échantillon est modélisée par la formule du théorème B. En supposant que notre échantillon de 20 % du recensement de 1991 est un échantillon aléatoire simple, nous avons tiré de celui-ci un certain nombre de sous-échantillons aléatoires simples pour plusieurs fractions de sondage allant de 0,00001 à 0,1, afin d'obtenir une courbe de Q_n en fonction de n. Par exemple, en tirant un sous-échantillon aléatoire simple avec une fraction de sondage de 50 % de l'échantillon de 20 %, on obtient un échantillon aléatoire simple de 10 % de la population canadienne. Pour chacun des sous-échantillons sélectionnés, nous avons calculé la proportion d'éléments uniques dans l'échantillon. Nous avons ensuite calculé la moyenne, c'est-à-dire Q_n, pour chacune des fractions de sondage. Avec cette information, nous avons estimé les paramètres α et β avec la méthode des moindres carrés. Les valeurs obtenues pour α et β sont respectivement 0.0891346 et

0.0000391. La valeur estimée de γ est donc de 0.000945. L'estimation de la probabilité conditionnelle avec un tel contenu pour le fichier de microdonnées des particuliers en 1991, c'est-à-dire lorsque la fraction de sondage est de 3 %, devient 0,73. Donc, pour ce contenu de 14 variables et une fraction de sondage de 3 %, les estimations obtenues sont un peu différentes : soit de 66 % avec la première approche et de 73 % avec la deuxième. Malgré cette différence, il en demeure que la probabilité conditionnelle lorsqu'on a une fraction de sondage de 3 % et un contenu de 14 variables est très élevée.

Pour établir un lien entre le calcul de la vraie probabilité conditionnelle avec les 9 variables recensées et l'estimation obtenue avec le contenu de 14 variables, nous avons estimé la valeur de P avec le contenu de 9 variables en utilisant la première méthode d'estimation décrite précédemment. À partir de l'échantillon du recensement, nous avons calculé les u_i du tableau croisé des 9 variables et obtenu la valeur de P = 0,433 pour une fraction de sondage de 20 %. Après avoir calculé la proportion d'uniques dans l'échantillon (0,07), on a estimé la proportion d'éléments uniques dans la population avec un contenu de 9 variables : 0,031 (= 0,433 x 0,07). Ainsi, pour ce contenu de 9 variables, on obtient une estimation de 3.1 % pour la proportion d'uniques dans la population. Notre estimation est donc très proche de la vraie valeur de 2.9 % calculée à partir de toute la population. Nous avons déterminé par la suite la proportion d'éléments uniques dans le fichier de microdonnées pour obtenir une estimation de la probabilité conditionnelle. Nous obtenons une probabilité conditionnelle de 16,6 %, qui est très proche de la vraie valeur de 15 % trouvée antérieurement.

Les résultats laissent voir que la probabilité conditionnelle, qui est le facteur central du risque de divulgation, dépend beaucoup plus du contenu de la population que de la fraction de sondage. En effet, l'ajout des cinq variables du questionnaire long du recensement a eu un impact très important sur cette probabilité. Ainsi, pour une fraction de sondage de 3 %, on a passé d'une probabilité conditionnelle de 15 % avec un contenu de 9 variables à une estimation de cette probabilité d'environ 70 % avec le contenu de 14 variables. Mais il faut toujours relativiser ces résultats à la possibilité qu'il existe des fichiers contenant ces variables.

3. Méthodes de réduction du risque de divulgation

3.1. Échange de données (data swapping)

L'échange de données consiste à échanger les valeurs de certaines variables entre les enregistrements. L'argument clé de cette technique est de créer un nombre restreint d'unités "artificielles" dans le fichier. L'introduction de ces unités rend impossible la

certitude absolue de faire des liens entre les enregistrements et les unités répondantes. Il faut bien voir cependant que cette méthode ne résoudra pas tous les problèmes. Il peut être aussi dommageable pour une unité répondante comme pour l'agence qu'un intrus affirme avoir obtenu une identification, même si cette dernière est en réalité fausse. Il faut en plus remarquer que, contrairement aux regroupements et suppressions de données, l'échange est par définition invisible à l'intérieur des données. En conséquence, un fichier de microdonnées pour lequel 1) on a opéré seulement un échange de données, et 2) on a laissé un contenu très détaillé, peut donner la fausse impression que l'agence ne fait à peu près rien pour sécuriser les réponses des unités ; ce que l'agence se doit d'éviter. Le propos de cette sous-section est de trouver l'impact de cette technique sur les estimations. Ceci est très important car cette technique n'est bonne que si le nombre d'enregistrements artificiels est relativement élevé. Nous allons trouver l'impact sur les estimateurs de totaux. Nous supposons que le fichier est auto-pondéré (facteur de pondération unique, noté W). Nous pouvons cependant faire beaucoup plus[10].

Nous avons un fichier de microdonnées $\mathscr{F}$ ayant m variables et n enregistrements. Nous allons le représenter par la matrice suivante :

$$\mathscr{F} = \left(Y_j(i) \right)_{j=1,\ldots,m\,;\; i=1,\ldots,n} .$$

Le symbole $Y_j(i)$ représente la valeur de la variable Y_j pour l'enregistrement i. Nous supposons que toutes les variables sont discrètes. Soit $\mathscr{F}$ un fichier de microdonnées. Un échange de données est la spécification de deux objets $\mathscr{E} = \{\wp, \sigma\}$ où 1) $\wp$ est une partition de l'ensemble des variables de $\mathscr{F}$ en deux parties A et B ; et 2) σ est une permutation de l'ensemble $[n] = \{1, \ldots, n\}$. Le résultat d'un échange de données est un fichier de microdonnées $\mathscr{F}^{\mathscr{E}}$ représenté par la matrice suivante :

$$\mathscr{F}^{\mathscr{E}} = \left(Y_j^{\mathscr{E}}(i) \right)_{j,i} = \begin{cases} Y_j(i) & si\, Y_j \in B \\ Y_{j(\sigma(i))} & si\, Y_j \in A \end{cases}$$

Le fichier de microdonnées $\mathscr{F}^{\mathscr{E}}$ est celui qui est publié, et par conséquent toutes les estimations et analyses seront faites à partir de ce dernier et non à partir de $\mathscr{F}$.

Une variable est dite permutable si elle est un élément de A, elle est dite fixe autrement. Le choix de la partition $\wp$ est crucial pour l'efficacité de l'échange de données. Nous pouvons voir que l'erreur causée par l'échange de données intervient

10. Boudreau, J-R. Impact d'un échange de données sur les estimateurs usuels. Rapport interne. Statistique Canada. 1994.

seulement lorsque la formule d'un estimateur utilise au moins une variable de A et de B. C'est-à-dire si toutes les variables utilisées pour une estimation se trouvent dans A ou dans B, l'échange de données n'a aucun effet. Par conséquent, si la relation entre deux variables est importante aux objectifs d'une enquête, ces deux variables devront être en même temps permutables ou fixes. Le support de $\mathscr{E}$, noté supp $\mathscr{E}$, est l'ensemble des enregistrements qui ne sont pas fixes par rapport à la permutation de $\mathscr{E}$. Par abus de langage, nous dirons également que supp $\mathscr{E}$ est le support de σ et quelques fois nous le noterons par supp σ. Pour un sous-ensemble D de [n], notons par n_D ou par #(D) la cardinalité de l'ensemble D. Nous définissons le taux de permutation de $\mathscr{E}$, noté τ, par le rapport de la cardinalité du support d'un échange sur n.

Nous allons maintenant définir la notion d'un domaine d'estimation. Soient $Y_1, \ldots, Y_r$ des variables de $\mathscr{F}$. Soit M_j un sous-ensemble de valeurs de la variable Y_j ($j = 1, \ldots, r$). Le domaine D défini par les variables Y_j et les valeurs M_j ($j = 1, \ldots, r$) est le sous-ensemble de [n] suivant :

$$D = D(Y_1, M_1; \ldots; Y_r, M_r) = \bigcap_{j=1}^{r} \bigcup_{s \in M_j} \{i \in [n]: Y_j(i) = s\}.$$

Ces ensembles sont les domaines d'intérêt les plus usuels. S'il y a au moins une variable permutable qui définit D, alors $\mathscr{E}$ va modifier le domaine. Ce nouveau domaine, noté $D^{\mathscr{E}}$, est donné par

$$D^{\mathscr{E}} = D(Y_1^{\mathscr{E}}, M_1; \ldots; Y_r^{\mathscr{E}}, M_r).$$

Par exemple, si toutes les variables définissant D sont permutables, alors $D^{\mathscr{E}} = \{ i : \sigma(i) \in D \}$, et nous dirons que $D^{\mathscr{E}}$ est un déphasage de D. Un domaine est fixe si toutes les variables définissant D sont fixes. Un domaine D peut toujours s'écrire comme l'intersection de deux parties P et F ($D = P \cap F$) où $P^{\mathscr{E}}$ est un déphasage de P et F est un domaine fixe. Nous dirons que le domaine est trivial si P ou F est égal à [n]. Ainsi $\mathscr{E}$ va générer une erreur pour un estimateur si et seulement si ce dernier est calculé sur un domaine non trivial. Soit D un domaine d'estimation quelconque et $\mathscr{E} = \{\wp, \sigma\}$ un échange de données. En général $D^{\mathscr{E}} \neq D$. Écrivons l'ensemble $D \cup D^{\mathscr{E}}$ en trois parties disjointes :

$$D \cup D^{\mathscr{E}} = D \cap D^{\mathscr{E}} \cup D - D^{\mathscr{E}} \cup D^{\mathscr{E}} - D.$$

La première partie est la partie invariante de D, la deuxième et la troisième sont la partie sortante et entrante respectivement. Elles sont notées respectivement par $D_0 = D_0^{\mathscr{E}}$, $D_\uparrow = D_\uparrow^{\mathscr{E}}$ et $D_\downarrow = D_\downarrow^{\mathscr{E}}$.

Notons par Σ^n l'ensemble des permutations de [n] et par Σ_k^n le sous-ensemble de Σ^n ayant un support de cardinalité k. Posons par convention $\#(\Sigma_0^0) = 1$. Nous allons prendre l'espace probabilisé correspondant à l'expérience de choisir au hasard une permutation ayant un support de cardinalité k. Alors les probabilités sont définies par :

$$P_{\Sigma_k^n}(\sigma) = \frac{1}{\#(\Sigma_k^n)},$$

pour tout $\sigma \in \Sigma_k^n$ (bien entendu, il faut que n et k nous permettent d'avoir $\#(\Sigma_k^n) > 0$). Soit $\mathscr{E} = (\wp, \sigma)$ un échange de données et X un estimateur quelconque. L'impact d'utiliser $\mathscr{F}^{\mathscr{E}}$ au lieu de $\mathscr{F}$ est fonction de la différence entre la réalisation de X provenant de $\mathscr{F}$ et la réalisation de X provenant de $\mathscr{F}^{\mathscr{E}}$ (notée $X^{\mathscr{E}}$). En fait, nous devons mesurer une variation autour de la "vraie" valeur X. Nous quantifierons cette erreur en utilisant la statistique suivante :

$$EQM_{\Sigma_k^n}(X^{\mathscr{E}}) = \sqrt{E_{\Sigma_k^n}(X^{\mathscr{E}} - X)^2} = \sqrt{V_{\Sigma_k^n}(X^{\mathscr{E}}) + B^2_{\Sigma_k^n}(X^{\mathscr{E}})};$$

où $V_{\Sigma_k^n}(X^{\mathscr{E}}) = E_{\Sigma_k^n}((X^{\mathscr{E}} - E^2_{\Sigma_k^n}(X^{\mathscr{E}}))$ et $B_{\Sigma_k^n}(X^{\mathscr{E}}) = E_{\Sigma_k^n}(X^{\mathscr{E}}) - X$ sont la variance et le biais de $X^{\mathscr{E}}$.

Nous allons calculer le biais pour un total quelconque. Notons par e_k la somme des (k + 1) premiers termes du développement de e^x autour de l'origine évalué au point -1. Premièrement, nous avons le résultat suivant.

Théorème C. *Pour $0 \le k \le n$, nous avons:*

$$\#(\Sigma_k^n) = \frac{n!}{(n-k)!} e_k .$$

Démonstration. Simple exercice d'analyse combinatoire.

Soit D un domaine quelconque non vide et $\mathscr{E} = \{\wp, \sigma\}$ un échange de données pour lequel $\sigma \in \Sigma_k^n$. Évaluons $P_{\Sigma_k^n}(\sigma : \sigma(d) \notin D)$ où $d \in D$. Enlevons momentanément

d de [n]. Si nous contrôlons les enregistrements fixes dans D et son complément, il suffit de transférer d avec ou bien un enregistrement du complémentaire de D qui est resté fixe ou un enregistrement qui a "bougé". Dans le premier cas, il faut faire bouger k - 2 enregistrements parmi les n - 1 enregistrements disponibles ; et dans le deuxième cas, il faut faire bouger k - 1 enregistrements parmi les n - 1 enregistrements. La probabilité recherchée devient donc

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \sum_{i=0}^{k-1} \frac{\binom{n_D - 1}{i}\binom{n - n_D}{k-1-i}\binom{k-1-i}{1} \#(\Sigma_{k-1}^{k-1})}{\#(\Sigma_k^n)}$$
$$+ \sum_{i=0}^{k-2} \frac{\binom{n_D - 1}{i}\binom{n - n_D}{k-2-i}\binom{n-n_D-k+2+i}{1} \#(\Sigma_{k-2}^{k-2})}{\#(\Sigma_k^n)}.$$

Si nous évaluons cette expression, nous obtenons

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \frac{e_{k-1}}{e_k}\frac{(k-1)!(n-k)!(n-1)!}{n!(k-1)!(n-k)!}\left[(k-1) - \frac{(n_D-1)(k-1)}{(n-1)}\right]$$
$$+ \frac{e_{k-2}}{e_k}\frac{(k-2)!(n-k)!(n-1)!}{(n!(k-1)!(n-k+1)!}\left[(n-n_D-k+2) - \frac{(n_D-1)(k-2)}{(n-1)}\right].$$

Ce qui donne

$$P_{\Sigma_k^n}(\sigma(d) \notin D) = \frac{1}{n-1}\left[1 - \frac{n_D}{n}\right]\left(\frac{(k-1)e_{k-1} + e_{k-2}}{e_k}\right)$$
$$= \frac{k}{n-1}\left(1 - \frac{n_D}{n}\right) = \left(\frac{n}{n-1}\right)\tau(1 - \delta_D)$$

où $\delta_D = \frac{n_D}{n}$ est la densité du domaine D.

Remarquons premièrement que pour un domaine quelconque $\#(D^{\sigma}) - \#(D) = \#(D_{\downarrow}) - \#(D_{\uparrow})$. Alors si P et F sont les parties déphasée et fixe de D, et si nous posons pour $i \in P^c$ et pour $j \in P$:

$$X_i(\sigma) = \begin{cases} 1 \text{ si } \sigma(i) \in P, \\ 0 \text{ autrement} \end{cases} \text{ et } Y_j(\sigma) = \begin{cases} 1 \text{ si } \sigma(j) \notin P, \\ 0 \text{ autrement} \end{cases}$$

nous avons la relation

$$\#(D^{\mathscr{E}}) - \#(D) = \sum_{i \in P^c \cap F} X_i - \sum_{j \in P \cap F} Y_j .$$

D'après ce qui précède, nous obtenons, si d ∈ P et d' ∉ P :

$$B_{\Sigma_k^n}\left(\#(D^{\mathscr{E}}) - \#(D)\right) = \#(P^c \cap F)\, P_{\Sigma_k^n}\left(\sigma : \sigma(d') \in P\right)$$

$$- \#(D)\, P_{\Sigma_k^n}\left(\sigma : \sigma(d) \notin P\right) = W n \tau \frac{n}{n-1}\left(\delta_F\, \delta_P - \delta_D\right).$$

Le biais relatif est donné par la formule suivante :

$$\frac{B_{\Sigma_k^n}(W \#(D))}{W \#(D)} = \frac{n}{n-1}\tau\left(\frac{\delta_P\, \delta_F}{\delta_D} - 1\right).$$

La formule du biais relatif est très instructive. Il peut être extrêmement périlleux de choisir un taux de permutation élevé. Nous ne pouvons pas supposer que le terme des densités est toujours très près de l'unité. Ce résultat discrédite un peu la technique.

3.2 Suppressions locales

La deuxième section donnait des moyens pour évaluer le risque de divulgation. Encore une fois, les conditions suffisantes d'avoir un risque peu élevé sont : une petite fraction de sondage, une proportion d'éléments uniques dans l'échantillon pas trop élevée et une estimation de α pas trop petite. Dès qu'une de ces conditions n'est pas respectée, ou bien il faut échantillonner de nouveau pour réduire la fraction de sondage ou modifier le contenu. Toute modification de contenu sera appelée un traitement dans les données. On peut effectuer soit un traitement global ou un traitement local. Un traitement global est appliqué à tous les enregistrements, comme, par exemple, un regroupement de valeurs d'une des variables d'appariement. Un traitement local, par opposition, n'est appliqué qu'à une partie des enregistrements. Il y a plusieurs méthodes de traitements globaux ou locaux. Elles ont toutes leurs points forts ou faibles. Nous aimerions répondre à la question suivante : lorsqu'on privilégie un traitement local, quels sont les enregistrements qui devraient être traités pour optimiser le traitement ?

En théorie, le but du traitement est de réduire le nombre d'éléments uniques dans la population qui se trouvent dans l'échantillon. Donc si nous voulons optimiser le traitement, nous devons trouver un moyen d'identifier ces enregistrements et deuxièmement d'appliquer un traitement qui ne les rende plus uniques dans la population. Pour ce qui est du traitement, nous allons donner un algorithme qui rend sécuritaires les enregistrements sans grand traitement. Reste donc la question du choix des enregistrements à traiter. Puisque les éléments uniques dans la population sont nécessairement uniques dans l'échantillon, nous devons nous concentrer premièrement seulement sur les uniques dans l'échantillon. Mais cela ne suffit pas. Il faut être en mesure de pouvoir filtrer les éléments uniques dans la population de ceux qui ne sont uniques que dans l'échantillon. C'est ici que le concept de la multiplicité d'un enregistrement s'insère dans la pratique. Comment peut-on faire pour filtrer les uniques dans la population des autres ? Il faut arriver à pouvoir évaluer le "degré d'unicité" des enregistrements ; à pouvoir dire qu'un enregistrement est plus unique qu'un autre. Comment faire ? Nous avons observé que la plupart des éléments uniques dans la population sont également uniques dans la population pour un sous-ensemble restreint de variables d'appariement. Autrement dit, nous avons observé que l'attribut d'unicité dans la population dépend surtout d'une combinaison très rare de valeurs d'un petit nombre de variables d'appariement. Cela dit, si nous recherchons les uniques dans la population avec, par exemple, seulement trois variables d'appariement, peut-être certains éléments seront déjà classés comme uniques. En cherchant les uniques pour toutes les combinaisons de trois variables parmi le nombre de variables d'appariement et en additionnant, pour chaque élément, le nombre de fois que ce dernier est unique, on en vient à une notion quantitative d'unicité. Le nombre de fois qu'un élément est unique dans un tableau à trois dimensions est appelé la multiplicité de cet élément. Plus un élément a une multiplicité élevée, plus cet élément a un risque d'identification élevé. Que se passe-t-il lorsque nous n'avons qu'un échantillon ? Nous avons constaté que si nous calculons la multiplicité seulement avec l'échantillon, elle définit une partition de l'échantillon dont les différentes parties ont des proportions d'éléments uniques dans la population très différentes. Nous avons simulé un petit exemple pour montrer l'efficacité du filtre.

Nous avons pris un échantillon aléatoire simple avec une fraction de sondage de 0,009 d'une population de taille 781 825 éléments. Ce qui donne une taille d'échantillon de 7 037. Le fichier contient cinq variables d'appariement. Le nombre d'éléments uniques dans la population est 35 718 (4,5 %). Le nombre d'éléments uniques dans l'échantillon s'élève à 2 301 (32,7 %). Le nombre d'éléments dangereux (uniques dans la population qui se trouvent dans l'échantillon) s'élève à 321 (4,5 %). La probabilité conditionnelle s'établit à 14 %. Si nous choisissons au hasard parmi les éléments uniques dans l'échantillon, seulement 14 % de ces enregistrements (en moyenne) sont dangereux. Beaucoup d'enregistrements qui ne requièrent aucun traitement sont tout de même traités. Si nous calculons la multiplicité des enregistrements, nous obtenons le *tableau suivant* :

Tableau 3
Résultats de la simulation

Multiplicité	# éléments	# uniques	%
10	18	15	83,3
9	41	23	56,1
8	64	33	51,6
7	45	26	57,8
6	191	61	31,9
5	220	77	35,0
4	140	33	23,5
3	388	32	8,2
2	294	17	5,8
1	472	3	0,6
0	5 164	1	0,0
Total	7 037	321	4,5

Nous pouvons voir aisément que la partition créée par la multiplicité nous aide grandement à choisir les enregistrements à traiter. Par exemple, si nous décidons de traiter tous les enregistrements ayant une multiplicité supérieure à trois, nous éliminons 83,4 % (268 éléments) des enregistrements dangereux en ne traitant que 10,3 % des enregistrements, ce qui est plus performant que d'y aller au hasard. Nous avons essayé cette technique avec des fichiers de dix ou quinze variables d'appariement et, bien que le filtre ne soit pas aussi performant que celui présenté ci-dessus, les résultats sont quand même surprenants. La recherche se poursuit maintenant vers une détermination de la multiplicité minimale où un traitement serait nécessaire. Cette multiplicité, appelée "le seuil de singularité" indiquerait, si le pourcentage de traitement est trop élevé, qu'il faut envisager plutôt des mesures globales.

Maintenant que nous savons quels enregistrements il faut traiter, concentrons-nous sur un algorithme qui fait le moins de suppression. L'objectif de la suppression est de rendre la multiplicité des enregistrements au-dessous d'un seuil acceptable. Voici l'algorithme pour un enregistrement :

I. Déterminer la fréquence de la valeur de chaque variable qui se trouve dans au moins un tableau à trois dimensions qui a servi à calculer la multiplicité.

II. Choisir la variable donnant la plus petite fréquence. Les égalités sont résolues au hasard.

III. Supprimer la valeur de cette variable.

IV. Éliminer tous les tableaux à trois dimensions où la variable supprimée est présente. Soustraire ces tableaux de la multiplicité.

V. Si la multiplicité est toujours supérieure au seuil, refaire une itération. Sinon le traitement de l'enregistrement est terminé.

3.3. Traitement pour les variables réelles

Nous présentons dans cette sous-section une proposition de traitement afin de réduire le risque de divulgation de la diffusion de variables réelles comme des sources de revenu. Il devient de plus en plus dangereux de publier les sources de revenu à l'unité près. Si nous voulons que ces dernières ne puissent pas être prises comme variables d'appariement, il faut introduire un certain bruit dans les valeurs de ces quantités. Arrondir au plus proche millier est une façon d'introduire du bruit dans les données. Certaines méthodes plus élaborées assurent l'invariabilité de certaines statistiques. Nous donnons ici un algorithme (l'arrondissement semi-contrôlé) qui garantit entre autres l'invariance des moyennes et des variances pour des sous-groupes très fins de la population.

Nous avons r variables réelles, appelées, V_1, ... , V_r, dans un fichier de microdonnées. Le fichier de microdonnées contient N enregistrements. Nous nommons $x^{i,j}$ la valeur de la variable V_j pour l'enregistrement i, c'est-à-dire $x^{i,j} = V_j(i)$. Nous avons alors un tableau à deux dimensions dans le fichier. De plus, nous avons un ensemble de conditions sur les variables. En effet, nous avons toujours une "colonne" donnant le total pour toutes les variables (la variable indexée r représente le total des variables réelles) :

$$\sum_{i=1}^{r-1} x^{i,j} = x^{i,r}$$

pour i = 1, ... , N. Si les variables sont des sources de revenu, V_r est la variable du revenu total de l'enregistrement. Voici le problème : nous voulons perturber les valeurs des variables, c'est-à-dire nous voulons déterminer de nouvelles variables Y_j (j = 1, ... , r) (le bruit ajouté) telles que les nouvelles valeurs $z^{i,j} = x^{i,j} + y^{i,j}$ (les valeurs qui seraient observées dans le fichier) répondent aux exigences suivantes :

1. $\sum_{j=1}^{r-1} z^{i,j} = z^{i,r}$ pour i = 1, ... , N (additivité des variables pour chaque enregistrement) ;

2. $\left|y^{i,j}\right| \leq Cte$ où la constante est déterminée à l'avance ;

3. Si $x^{ij} = 0$ alors $z^{ij} = 0$ pour tout i, j (les valeurs nulles demeurent nulles) ;

4. $\sum_{i=1}^{N} l_A(i)\, z^{i,j} z^{i,j'} = \sum_{i=1}^{N} l_A(i)\, x^{i,j} x^{i,j'}$ pour j, j' = 1, ... , r et A est n'importe quel domaine d'estimation.

La dernière contrainte a trait à l'invariance des variances et covariances. Nous croyons qu'il n'existe pas de solution non triviale à ce problème, c'est-à-dire une avec une valeur y^{ij} non nulle pour au moins un couple d'indices. Cependant, nous pouvons rechercher des solutions partielles ou approximatives. L'arrondissement semi-contrôlé nous donne des solutions partielles en plus d'être peu coûteux.

Tout d'abord, nous disons qu'un tableau est arrondi si toutes ses entrées appartiennent à un idéal (Cte) où Cte est un entier strictement supérieur à l'unité. Bien entendu, la valeur absolue de la différence entre l'entrée et l'élément de l'idéal se doit d'être la plus petite possible. Une marginale d'un tableau arrondi est contrôlée si la somme des valeurs arrondies définissant la marginale est égale à la valeur arrondie de la marginale. L'idée maîtresse de l'arrondissement semi-contrôlé est de contrôler le grand total d'un tableau à deux dimensions et de laisser la propriété d'additivité du tableau s'occuper du contrôle des marginales du tableau. L'arrondissement est dit "semi-contrôlé" parce que le contrôle sur les marginales n'est pas parfait. Un petit exemple sera instructif. Supposons le tableau 2 x 2 suivant avec ses marginales :

2	5	7
3	6	9
5	11	16

On peut représenter ce tableau par la notation compacte suivante : (2, 5, 7 | 3, 6, 9 | 5, 11, 16). Posons Cte = 5. Alors le grand total peut être arrondi soit à 15, soit à 20. Supposons 15. L'algorithme trouve le tableau temporaire suivant : (0, 5, ? | 0, 5, ? | ?, ?, 15). Les valeurs sont les éléments de l'idéal (5) non supérieurs aux entrées. Les marginales sauf le grand total sont toutes à déterminer. Si nous faisons la somme de toutes les entrées à l'intérieur du tableau (éléments qui ne sont pas des marginales), nous obtenons 10. Puisque le grand total est établi à 15, nous devons additionner à

une entrée interne du tableau temporaire la valeur 5 pour donner également le grand total de 15. Le choix d'une entrée détermine l'arrondissement des éléments internes du tableau. Supposons que nous choisissons l'élément (2,1) du tableau. Nous avons le tableau temporaire (0, 5, ? | 5, 5, ? | ?, ?, 15). Rendu à cette étape, nous déterminons les marginales arrondies en sommant les éléments internes arrondis qui définissent les marginales. Le tableau final arrondi donne (0, 5, 5 | 5, 5, 10 | 5, 10, 15). Ce tableau est additif et parfaitement contrôlé. L'algorithme donne toujours un tableau additif mais pas nécessairement contrôlé. Un choix judicieux des éléments internes où il faut additionner 5 peut donner un contrôle presque parfait.

Nous donnons une solution partielle à notre problème : nos candidats pour $y^{i,j}$. Soit Cte > 0 une constante spécifiée à l'avance. Elle est appelée la base d'arrondissement. Voici l'algorithme :

1. Créer une partition très fine d'enregistrements (ceci afin de simuler un contrôle sur certains domaines populaires). Par exemple pour un fichier de particuliers, nous pourrions regrouper les enregistrements par l'âge, le genre de la personne, l'état matrimonial, etc... Supposons que cela donne M groupes $G_1, \ldots, G_M$.

2. Calculer le revenu total de chaque groupe : $T_m = \sum_{i \in G_m} x^{i,r}$.

3. Calculer le revenu total global : $T = \sum_{m=1}^{M} T_m$.

4. Arrondir d'une manière aléatoire et sans biais T en utilisant Cte comme base d'arrondissement. Nous l'appellerons T_a.

5. Pour m = 1, ... , M, Calculer $B_m = \left[\frac{T_m}{Cte}\right] \times Cte$, [a] est le plus grand entier inférieur ou égal à a.

6. Calculer $n = \frac{T_a - B_1 - \ldots - B_M}{Cte}$.

7. Trier les valeurs (T_m - B_m) par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers n groupes de la liste triée le nouveau total $T_{ma} = B_m + Cte$. Attribuer $T_{ma} = B_m$ pour les autres groupes.

Maintenant, nous allons suivre les étapes suivantes indépendamment pour chaque groupe. Nous fixons m.

8. Calculer $T_m^j = \sum_{i \in G_m} x^{i,j}$ pour j = 1, ... , r - 1

9. Calculer $B_m^j = \left[\frac{T_m^j}{Cte}\right] x\ Cte$,

10. Calculer $n_m = \frac{T_{ma} - B_m^1 - \ldots - B_m^{r-1}}{Cte}$.

11. Trier les valeurs (T_m^j - B_m^j) par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers nm colonnes ou variables sur la liste triée le nouveau total $T_{ma}^j = B_m^j + Cte$. Attribuer $T_{ma}^j = B_m^j$ pour les autres variables.

Maintenant, nous allons faire les étapes suivantes indépendamment pour chaque variable de chaque groupe. Nous fixons m et j. Ainsi les valeurs que nous considérons font partie du groupe G_m et de la variable V_j.

12. Calculer $B_m^{i,j} = \left[\frac{x^{i,j}}{Cte}\right] x\ Cte$ pour l'enregistrement i,

13. Calculer $n_m^j = \frac{T_{ma}^j - B_m^{1,j} - \ldots}{Cte}$.

14. Trier encore une fois les valeurs ($T_m^{i,j}$ - $B_m^{i,j}$) par ordre décroissant. L'ordre des valeurs égales est résolu aléatoirement. Assigner les premiers n_m^j enregistrements sur la liste triée la nouvelle valeur $z^{i,j} = B_m^{i,j} + Cte$. Attribuer $z^{i,j} = B_m^{i,j}$ pour les autres enregistrements.

15. Calculer $z^{i,r} = \sum_{j=1}^{r-1} z^{i,j}$.

L'algorithme d'arrondissement perturbe les variables réelles en utilisant des bases. Nous pouvons avoir plusieurs bases : une pour les petites valeurs, une pour les valeurs intermédiaires et une pour les grandes valeurs. Ces bases déterminent la quantité de bruit ajouté aux données. Nous devons choisir des bases assez grandes pour que la possibilité d'une ré-identification soit minimale ; et en même temps, les bases doivent être aussi basses que possible pour maintenir l'utilité des données. Nous nous retrouvons devant un problème classique d'optimisation. L'algorithme est construit pour maintenir la cohésion maximale des données étant données les bases d'arrondissement. Le problème revient alors à évaluer la possibilité d'une ré-identification. Normalement une identification survient lorsqu'il y a un vrai couplage biunivoque entre le fichier de microdonnées et un fichier contenant des

identificateurs uniques comme noms, adresses, etc... Nous observons une ré-identification lorsque qu'il y a encore un vrai couplage biunivoque bien que l'arrondissement est opéré avant le couplage. Bien entendu, nous devons considérer des algorithmes de couplages statistiques. Ces algorithmes dépendent de distances entre les enregistrements des deux fichiers. Nous pensons que seules les données doivent déterminer le choix de la distance. Nous donnons ici la distance que nous préconisons. Nous considérons le type général de fonctions de distance donné par la formule suivante :

$$D(l_1, l_2) = \sqrt{\sum_{i=1}^{r} w_i(v_i(l_1) - v_i(l_2))^2}.$$

où w_i est un facteur de pondération associé à la variable i (1 = 1, ... , r) et $v_i(l)$ est la valeur de la variable i pour l'enregistrement l. Nous considérons seulement des vecteurs $w = (w_1, \ldots, w_r)$ de pondération normalisés, c'est-à-dire tels que $w_i \geq 0$ (1 = 1, ... , r) et la somme des composantes est l'unité. Étant donné w, on peut trouver, pour chaque enregistrement non arrondi, l'enregistrement arrondi le plus près d'après la fonction de distance choisie. Il est alors possible de trouver la proportion d'enregistrements non arrondis pour lesquels leur enregistrement arrondi le plus proche associé est le même enregistrement. Soit P_w cette proportion. Soit P le maximum des P_w où w parcourt un sous-ensemble dense de son domaine de définition et soit w^* le vecteur des facteurs de pondération qui donne le maximum. Il faut utiliser cette fonction de distance pour évaluer la possibilité de ré-identifications. Ainsi, en spécifiant une proportion de ré-identifications acceptable, nous pouvons trouver les bases d'arrondissement.

4. Conclusion

Nous avons décrit dans cet article comment nous avons abordé le problème d'assurer la confidentialité des réponses recueillies pour quelques enquêtes de Statistique Canada. La recherche se poursuit sur plusieurs fronts. Premièrement, nous recherchons des méthodes d'estimation de la probabilité conditionnelle et plus généralement de l'estimation du nombre d'éléments uniques dans une population qui soient plus performantes. Nous jugeons que cela est essentiel pour arriver à une bonne évaluation du risque de divulgation. Nous essayons aussi de trouver une justification du concept de la multiplicité d'un échantillon. Peut-être arriverons-nous à une détermination du seuil de singularité. Nous regardons également si nous pouvons améliorer l'arrondissement semi-contrôlé. Toutes ces recherches sont nécessaires car les pressions pour plus d'information sont plus vives que jamais. L'ère de l'information n'est pas près de s'éclipser.

LA PRATIQUE DES ENQUÊTES PAR TÉLÉPHONE À STATISTIQUE CANADA

Jean-François Gosselin

1. INTRODUCTION

La pratique des enquêtes téléphoniques occupe présentement une place prépondérante parmi les activités de collecte de données de Statistique Canada. En fait, au cours des vingt-cinq dernières années, la pratique des enquêtes par téléphone n'a cessé de prendre de l'ampleur. Par exemple, nous avons observé une augmentation graduelle du nombre d'interviews téléphoniques de l'Enquête sur la population active (EPA) depuis le début des années 70 ; auparavant, toutes les interviews se faisaient sur place. A l'heure actuelle, environ 85 % des interviews de l'EPA sont effectuées par téléphone. Par ailleurs, la collecte des données du programme des enquêtes agricoles a été complètement remaniée au cours des années 80, ce qui a mené à l'abandon des enquêtes postales et à la mise sur pied d'un nombre important d'enquêtes téléphoniques. Enfin, la méthode d'interview téléphonique s'est avérée très pratique pour répondre aux besoins particuliers et très variés de notre programme des enquêtes spéciales.

Plusieurs raisons ont motivé ces changements. Dans le cas de l'EPA, un programme de recherche fructueux a permis de fonder les changements sur des bases méthodologiques et statistiques solides et de déterminer les risques et les limites. La nécessité de trouver des moyens de réduire les coûts de collecte tout en maintenant la qualité a toujours été et demeure un souci constant de Statistique Canada. Bien sûr, la conjoncture actuelle oblige les organismes de statistique à trouver des solutions de rechange à la fois souples, efficaces et d'exécution rapide pour remplacer les interviews directes très coûteuses.

Par ailleurs, comme dans le cas des enquêtes-entreprises, les méthodes téléphoniques sont un complément essentiel aux enquêtes postales qui, malgré leur coût avantageux, produisent des taux de réponse décevants.

Fait important à noter, les enquêtes téléphoniques se prêtent très bien aux méthodes d'interview assistée par ordinateur qui facilitent grandement la gestion des cas et ouvrent la voie à l'utilisation de questionnaires très complexes. Nous avons observé non seulement une nette évolution au niveau de la complexité, mais aussi une nette

progression en ce qui a trait à la durée des interviews et à la nature délicate des sujets traités.

Nous présenterons d'abord une brève description de la structure opérationnelle de Statistique Canada ainsi qu'une vue d'ensemble du programme d'enquêtes et des pratiques courantes. Nous situerons ensuite l'évolution des méthodes de collecte de certaines enquêtes dans un contexte historique. En guise de conclusion, nous présenterons quelques perspectives d'avenir.

2. ORGANISATION DES OPÉRATIONS DE STATISTIQUE CANADA

2.1 Les opérations régionales

Pour effectuer ses enquêtes, Statistique Canada dispose d'un réseau de bureaux régionaux. Les cinq premiers bureaux ont été créés en 1945 ; situés à Halifax, Montréal, Toronto, Winnipeg et Vancouver, ces bureaux ont d'abord été ouverts pour les besoins de l'Enquête sur la population active, laquelle a récemment été remaniée. Mais on prévoyait déjà qu'ils pourraient servir à de nombreuses autres enquêtes ; d'ailleurs, peu de temps après, un programme d'enquêtes sur les prix a été proposé et un certain nombre d'enquêtes sur les ménages ont été établies. Aujourd'hui, des bureaux existent aussi à Sturgeon Falls et à Edmonton.

Au cours des années 80, l'installation de nouveaux mini-ordinateurs a permis de faire l'essai de nombreux projets qui ont élargi davantage le champ des activités régionales. L'un de ces projets a permis d'intégrer avec succès la méthode d'interview téléphonique assistée par ordinateur (ITAO) à l'Enquête mensuelle sur les industries es manufacturières en 1987.

Dans les années 90, les régions ont connu une véritable révolution technologique avec l'intégration à grande échelle des interviews assistées par ordinateur (IAO) aux enquêtes sur les ménages, l'agriculture et les entreprises. Les interviews sur place assistées par ordinateur (IPAO) sont utilisées pour l'Enquête sur la population a active et des enquêtes supplémentaires ou longitudinales sur les ménages, qui traitent d'un large éventail de sujets tels que la dynamique du travail et du revenu, la santé et les enfants.

Les opérations régionales sont responsables des opérations des programmes suivants :

- **Programme des enquêtes-ménages**, y compris l'Enquête mensuelle sur la population active, les enquêtes supplémentaires (équipement ménager, finances des consommateurs, enquêtes spéciales), les enquêtes sur les dépenses des consommateurs et les enquêtes longitudinales.

- **Programme des enquêtes-entreprises**, y compris les enquêtes mensuelles sur les secteurs manufacturiers, du commerce de gros et de détail, et les enquêtes annuelles sur les secteurs du commerce de gros et de détail, et des services.

- Les relevés mensuels pour **l'Indice des prix à la consommation**.

- Les activités reliées à la mise à jour du **Registre des entreprises**.

- Un programme **d'enquêtes spéciales** auprès des ménages et des entreprises, dont le principal mode de collecte est l'interview téléphonique assistée par ordinateur.

- D'autres activités de collecte reliées aux permis de construction, aux activités culturelles et au camionnage.

2.2 Opérations du bureau central

Comparativement aux opérations régionales, l'histoire des opérations du bureau central est beaucoup plus récente. Avant 1985, le personnel de la plupart des divisions des programmes assurait le soutien opérationnel des activités statistiques de leur programme respectif. C'était le cas, par exemple, des programmes des industries manufacturières, du commerce de gros, du commerce de détail, des services, de la construction, des transports, de la santé, de l'éducation, de la culture, des établissements publics, etc. Dès s le début de 1985, Statistique Canada a entrepris ce qu'on a appelé l'intégration des opérations" du bureau central, et le personnel a été progressivement "intégré" à une seule division.

L'idée d'intégrer les opérations du bureau central circulait depuis plusieurs années. Certains étaient d'avis que l'on pouvait réaliser des progrès notables en regroupant les opérations et en libérant certaines ressources des contraintes de la gestion opérationnelle, pour qu'elles puissent se consacrer aux produits, aux services et aux programmes eux-mêmes. Cette conviction est devenue réalité dans les années 80, lorsque la haute direction de Statistique Canada a voulu consolider et intensifier la fonction d'analyse tout en décuplant le rendement malgré de profondes compressions budgétaires. Pour relever le défi, il fallait donc rassembler les ressources des opérations ions du bureau central qui étaient dispersées dans un grand nombre de divisions différentes. C'est dans un tel contexte qu'a germé, en 1984, l'idée d'intégrer les opérations du bureau central et de donner à celui-ci la

responsabilité principale des grandes enquêtes annuelles qui exigent énormément de vérification et de traitement.

L'intégration des opérations a permis d'accroître grandement l'efficience. On a pu réduire de 25 % le nombre d'années-personnes du secteur des opérations, sans pour autant sacrifier la qualité des programmes, mais en diminuant les frais généraux et en tirant le meilleur parti possible de certaines initiatives. Par exemple, on a supprimé les échelons intermédiaires en élargissant le rapport entre la surveillance et le nombre d'employés ployés à surveiller. On a également utilisé plus efficacement le potentiel en ressources des opérations en étalant les périodes de pointe et les périodes creuses dans le calendrier.

Toutefois, on n'aurait jamais pu réaliser des économies d'une telle envergure si le concept d'une gestion des opérations mieux disciplinée n'avait pas été renforcé par un programme vigoureux de recherche et de développement. Ce programme a été mis en place afin de concentrer et d'orienter les efforts pour atteindre l'objectif visé : professionnaliser les opérations. Il comprend l'analyse des opérations et la conception des tâches, la restructuration, la formation du personnel, le contrôle et l'assurance de la qualité, et l'automatisation des procédés. Par exemple, des travaux de recherche sur les méthodes d'ITAO ont mené à la mise sur pied d'une unité spéciale responsable des enquêtes téléphoniques.

Les opérations centrales sont responsables des opérations des programmes suivants :

- Les **grandes enquêtes annuelles** dans les secteurs manufacturiers, du commerce de gros et de détail, des services, des transports, ainsi que les enquêtes trimestrielles et annuelles sur les finances des entreprises.

- La collecte de données auprès des **établissements** de santé, d'enseignement, culturels ou publics. Plusieurs de ces programmes font grand usage de données administratives accessibles sur support informatique.

- Une unité chargée de mener des **enquêtes spéciales** à partir d'un endroit central à Ottawa, selon la méthode d'ITAO.

- Une multitude d'activités reliées au programme de voyages internationaux, au codage des données de l'EPA, à la cartographie, à la saisie de données, etc.

Aujourd'hui, la Division des opérations et de l'intégration est responsable d'un très grand nombre d'opérations différentes et compte 300 employés.

2.3 Évolution du mode de collecte : un portrait global

Globalement, nous avons observé des changements importants en ce qui a trait aux modes de collecte utilisés au cours des cinq dernières années. Par exemple, le *tableau 1* présente le nombre de répondants selon le mode de collecte depuis 1991 pour les opérations de collecte dans les bureaux régionaux.

Plusieurs observations se dégagent de ce tableau. Premièrement, on note une diminution importante du nombre d'unités déclarantes. Cette baisse d'environ un demi-million d'unités, soit tout près de 20 %, est principalement le résultat d'efforts considérables pour diminuer le fardeau des répondants des enquêtes-entreprises. En particulier, l'Enquête sur la rémunération et les heures de travail (ERHT) a été remaniée en profondeur. Une utilisation judicieuse des données fiscales, accompagnée de la création d'une enquête d'appoint de taille réduite appelée Enquête mensuelle sur la rémunération (EMR) auprès des entreprises, a permis de réduire de façon remarquable le fardeau de réponse et les coûts liés à l'ERHT. On note également une légère augmentation du nombre d'unités déclarantes chez les agriculteurs, alors que les variations observées pour les enquêtes-ménages ne sont pas inhabituelles et reflètent principalement la taille du programme des enquêtes spéciales.

Deuxième fait important, les questionnaires postaux ne sont plus du tout utilisés pour les enquêtes-ménages et agricoles, tandis que cette méthode est encore la préférée pour les enquêtes-entreprises, et ce, pour des raisons d'économie. Cependant, il y a eu une diminution importante de l'utilisation de ce mode de collecte, soit environ 50 %, ce qui représente le double de la baisse correspondante du nombre d'entreprises.

Quant au mode d'interview téléphonique, il est passé globalement de 65 % à 75 % au cours des cinq dernières années. Il s'est maintenu à de très hauts niveaux pour les enquêtes-ménages (81 % à 87 %) et agricoles (89 % à 94 %), mais il a fait un bond de 45 % à 63 % pour les enquêtes-entreprises. En 1991, le ratio du nombre d'unités déclarantes par la poste au nombre d'unités déclarantes par téléphone était de 1:1 ; en 1996, il est passé à 1:2, ce qui démontre une nette progression du mode d'interview téléphonique pour les enquêtes-entreprises.

Nous discuterons maintenant d'exemples particuliers de programmes qui ont connu une évolution marquée en ce qui a trait à l'utilisation du mode de collecte par téléphone, en commençant par l'EPA.

Tableau 1
Nombre de répondants selon le mode de collecte
Opérations régionales

SECTEUR	ANNÉE	MODE DE COLLECTE				TOTAL
		POSTE	Interview en personne	Interview au téléphone #	Interview au téléphone %	
MÉNAGES	91/92	-	180 560	1 187 040	86,79	1 367 600
	92/93	103 370	195 200	1 324 820	81,60	1 623 390
	93/94	800	198 222	871 920	81,41	1 070 942
	94/95	-	269 007	1 134 638	80,83	1 403 645
	95/96	-	173 619	964 856	84,74	1 138 475
ENTREPRISES	91/92	763 460	107 300	723 360	45,36	1 594 120
	92/93	651 130	96 640	784 270	51,19	1 532 040
	93/94	624 824	85 300	782 377	52,42	1 492 501
	94/95	409 405	85 720	771 293	60,90	1 266 418
	95/96	411 464	40 975	759 652	62,67	1 212 091
AGRICULTURE	91/92	-	12 700	153 230	92,34	165 930
	92/93	-	10 400	169 750	94,22	180 150
	93/94	-	24 380	207 708	89,49	232 088
	94/95	-	12 845	197 183	93,88	210 028
	95/96	-	12 848	199 974	93,96	212 822
INSTITUTIONS PUBLIQUES	91/92	37 980	140	27 550	41,95	65 670
	92/93	37 980	140	27 550	41,95	65 670
	93/94	36 100	252	28 500	43,94	64 852
	94/95	35 600	-	29 000	44,89	64 600
	95/96	2 338	-	12 000	83,69	14 338
TOTAL	91/92	801 440	300 050	2 091 180	65,48	3 193 320
	92/93	792 480	295 980	2 306 390	67,81	3 401 250
	93/94	661 724	178 879	1 890 505	66,09	2 860 383
	94/95	445 005	150 152	2 132 114	72,40	2 944 691
	95/96	413 802	69 656	1 936 482	75,12	2 577 726

3. L'ENQUÊTE SUR LA POPULATION ACTIVE

Avec son échantillon mensuel de 56 000 ménages, l'Enquête sur la population active du Canada est de loin la plus importante enquête-ménage menée par Statistique Canada. Elle utilise une base de sondage aréolaire à plusieurs degrés et un échantillon avec renouvellement, c'est-à-dire que les ménages demeurent dans l'échantillon pendant six mois consécutifs avant d'en être supprimés. Les données sont recueillies chaque mois par un personnel de 1 000 intervieweurs répartis dans tout le Canada.

Jusqu'au début des années 70, toutes les interviews étaient faites sur place, tandis qu'aujourd'hui environ 85 % se font par téléphone. Cela est le résultat d'un programme de recherche fructueux qui a permis de fonder de tels changements sur des bases méthodologiques et statistiques solides et de déterminer les risques et les limites.

Quelles ont été les étapes importantes de ce changement en profondeur ? C'est ce que nous allons maintenant voir. Pour plus de détails, voir Drew (1991) qui présente une excellente description des activités de recherche dans ce domaine.

3.1 Une première étape

C'est au début des années 70 que la méthode d'interview téléphonique a été employée pour la première fois dans le cadre de l'EPA. Les raisons qui ont motivé ce choix étaient principalement le désir de réduire les coûts et les délais de publication. Or, un tel changement en profondeur n'a pu se faire sans effectuer d'abord des tests permettant de démontrer que l'impact sur les taux de réponse et la qualité des données serait négligeable.

Un test a été effectué à Toronto et à Vancouver au début de 1971 pour démontrer la faisabilité de recueillir par téléphone les données de l'EPA pour toutes les interviews, sauf celles du premier mois. C'est ce qu'on appelle une "enquête téléphonique à chaud", par opposition à une "enquête téléphonique à froid", c'est-à-dire une interview téléphonique qui n'est pas précédée d'une interview sur place (Groves et coll. 1988).

En 1972 et 1973, le test a été étendu aux autres grands centres métropolitains. Une comparaison entre la nouvelle méthode et un groupe de contrôle (Muirhead et coll. 1975) a clairement démontré qu'il était possible de réduire les coûts de collecte d'environ 17 % sans nuire aux taux de réponse. De plus, aucune différence importante n 'a été décelée en ce qui a trait aux taux de participation à la population active et de chômage. Sur le plan opérationnel, cela nous a permis d'augmenter la productivité des intervieweurs ; au lieu d'interviewer de 45 à 55 ménages, ils pouvaient maintenant en interviewer de 70 à 90. Le taux de collecte par téléphone s'est stabilisé à environ 74 %, comparativement à une valeur théorique maximale de 83 % correspondant au 5/6 de l'échantillon. La différence est attribuable à divers problèmes opérationnels comme la non-disponibilité d'un téléphone. Seulement 1,2 % des ménages ont refusé de fournir l'information par téléphone, ce qui a démontré un grand niveau d'acceptabilité de la part des répondants.

À cette époque, le mode d'interview téléphonique a été limité aux principales régions urbaines en raison des préoccupations entourant la confidentialité des données et de la forte incidence des lignes partagées dans certaines régions urbaines

et rurales. Or, dans le cadre du remaniement de l'EPA de 1981, un test (Choudhry 1984) a produit des résultats similaires à ceux obtenus dans les grandes régions urbaines, démontrant ainsi une acceptation de la part des répondants et la faisabilité d'effectuer l'interview téléphonique à chaud dans toutes les régions. La méthode a donc été mise en place à partir de 1984, ce qui a permis de réduire les coûts de collecte de 10 %.

Ces expériences très positives nous ont alors amené à soulever des questions supplémentaires. Jusqu'où peut-on pousser l'utilisation de l'interview téléphonique ? Peut-on employer la méthode à froid pour l'EPA ? Est-il possible de remplacer la base aréolaire par une base de sondage téléphonique ? La méthode d'interview assistée par ordinateur (IAO) peut-elle être mise à profit pour l'EPA et les autres enquêtes-ménages ? Drew (1991), dans son article sur la recherche pour les méthodes d'enquêtes par téléphone, traite de ces questions de façon détaillée. Nous présenterons ici un résumé des principaux résultats.

3.2 L'interview téléphonique à froid

La méthode **d'interview téléphonique à froid avec suivi sur place** a été mise à l'essai au cours de la période de 1985 à 1989. Les logements nouvellement échantillonnés de l'EPA ont été appariés, à partir de l'adresse, aux listes obtenues des compagnies de téléphone. Le taux d'appariement observé a été de 65 %. Des échantillons d'essai et de contrôle ont alors été construits de manière à faire une évaluation statistique de la méthode traditionnelle et d'interview à froid. Les intervieweurs devaient mener une interview téléphonique auprès des logements compris dans le test et faire un suivi sur place seulement si cela était nécessaire. Même si aucune différence significative n'a été relevée à l'égard des taux de réponse, l'échantillon pour essai a permis de déceler une sous-estimation de la population active et du taux de chômage pour certaines catégories de population au Québec. Une analyse plus approfondie a révélé l'existence d'un lien avec un programme provincial d'inspection des bénéficiaires d'aide sociale. L'interview téléphonique à froid semble donc être plus sensible à des situations exogènes à l'enquête, comme celle mentionnée.

On a également mis à l'essai la méthode **d'interview à froid sans suivi sur place** avec et sans lettre de présentation. Même si la lettre de présentation a eu un effet positif sur les taux de réponse, une comparaison des deux méthodes a révélé une augmentation importante de la non-réponse par rapport au groupe contrôle. Ce résultat est significatif car une analyse semble indiquer la présence d'un biais important dû à la non-réponse à l'égard du taux de chômage.

Compte tenu de ces résultats, il a été décidé de ne pas adopter la méthode d'interview téléphonique à froid de façon générale pour l'EPA. Cependant, la

méthode a été mise à profit dans deux cas bien précis. Dans les immeubles à étages multiples où les taux de réponse étaient insatisfaisants à cause de problèmes d'accès, l'utilisation de cette méthode a permis de réduire l'écart dans la non-réponse de façon considérable. Un autre changement important au mode de collecte a été le suivi téléphonique auprès des logements de l'échantillon qui n'ont pu être contactés durant le premier mois de l'enquête. Il est important de noter qu'une lettre de présentation a été utilisée dans tous les cas.
L'effet global de toutes ces mesures a fait grimper le taux d'interview téléphonique pour l'enquête à 80 % en 1985 et à 83 % en 1990.

3.3 Base de sondage téléphonique

Au Canada, la couverture téléphonique est excellente. En effet, la situation n'a cessé de s'améliorer depuis les années 70 et elle se situe depuis dix ans à environ 98,5 % (voir *tableau 2*).

Tableau 2
Ménages sans téléphone par province (%)

	1976	1981	1985	1987	1990	1996
Terre-Neuve	10,1	6,5	5,6	3,5	2,2	2,6
Île-du-Prince-Édouard	-	5,3	5,0	4,5	2,3	2,0
Nouvelle-Écosse	7,5	5,1	3,7	3,2	1,5	1,7
Nouveau-Brunswick	5,7	5,8	5,1	3,2	2,3	1,8
Québec	3,3	2,3	1,5	1,7	1,5	1,2
Ontario	2,5	1,9	1,0	1,0	1,2	1,2
Manitoba	4,1	2,2	2,8	2,0	1,9	1,7
Saskatchewan	3,6	2,4	2,7	2,4	2,4	1,6
Alberta	2,9	2,6	2,1	1,9	2,1	1,0
Colombie-Britannique	4,1	2,9	2,4	1,4	1,5	1,3
Canada	3,5	2,5	1,8	1,6	1,5	1,3

Source : Statistique Canada, estimations provenant de l'enquête sur les installations et l'équipement ménager.

Laflamme (1990) a réalisé une étude qui démontre que l'incidence des numéros de téléphone non publiés peut atteindre 10 % et plus au niveau provincial. Il a aussi comparé les caractéristiques des personnes vivant dans un ménage dont le numéro est non publié ou qui n'ont pas de téléphone.
Ces résultats sont reproduits au *tableau 3* (Drew 1991).

Tableau 3
Caractéristiques de la population active selon le statut téléphonique

Province	Statut téléphonique	Taux de chômage	Taux d'activité
Nouvelle-Écosse	Publiés	9,0	71,9
	Non publiés	9,8	70,2
	Pas de téléphone	17,2	62,3
Alberta	Publiés	6,3	80,7
	Non publiés	8,2	81,5
	Pas de téléphone	11,1	67,0

Comme nous pouvons le constater, les caractéristiques de la population active diffèrent de façon considérable, plus particulièrement pour les personnes sans téléphone. Bien que cette dernière catégorie ne représente que 1,5 % de la population, la possibilité d'utiliser une base de sondage qui ne tient pas compte des personnes sans téléphone n'est pas une option valable, étant donné l'importance de l'enquête et la précision requise dans les estimations de l'emploi et du chômage. Par ailleurs, la technique de composition téléphonique aléatoire (CTA), même si elle tient compte des numéros non publiés, entraîne des difficultés importantes de mise à jour dans le cas des enquêtes par panel.

Finalement, les raisons qui justifient le maintien de la base de sondage aréolaire sont non seulement d'ordre méthodologique, mais aussi d'ordre stratégique. En effet, le réseau d'intervieweurs de l'EPA est largement mis à profit pour plusieurs autres enquêtes qui exigent un personnel compétent sur le terrain. Le démantèlement de ce précieux réseau affaiblirait notre infrastructure et diminuerait ainsi notre capacité de réaliser certains types d'enquêtes. Pour ces raisons, nous avons opté pour le maintien de la base aréolaire.

3.4 L'interview assistée par ordinateur

Des études préliminaires (Catlin et coll. 1988) portant sur l'interview assistée par ordinateur (IAO) ont démontré des améliorations quantitatives marquées , soit une diminution importante des rejets à la vérification et une élimination des erreurs de branchement.

4. LES ENQUÊTE-MÉNAGES TÉLÉPHONIQUES

Traditionnellement, le véhicule par excellence pour répondre aux besoins des enquêtes spéciales a été l'enquête supplémentaire à l'EPA. Celle-ci pouvait s'appuyer

sur une infrastructure comportant une base de sondage permanente et un réseau d'intervieweurs qualifiés, répartis dans tout le Canada, capables de répondre à des besoins particuliers. Cependant, au cours des dix dernières années, nous avons observé une popularité croissante des enquêtes téléphoniques.

Plusieurs facteurs peuvent expliquer ce phénomène :

- La technique de composition téléphonique aléatoire (CTA) et l'interview téléphonique à froid ne semblaient pas convenir à l'EPA pour les raisons mentionnées précédemment ; par contre, ces outils se sont révélés très utiles dans le contexte des enquêtes téléphoniques de taille moyenne, dans les cas où il n'existait pas d e base de sondage suffisamment complète, et pour lesquelles l'impact du problème de non-couverture des ménages sans téléphone était jugé non significatif. En particulier, l'Enquête sociale générale (ESG) a été mise en oeuvre en 1985 (taille d'échantillon de 10 000 ménages) pour combler des besoins statistiques précis. Nous en sommes maintenant au 11e cycle de l'ESG ; les cycles de l'enquête ont porté sur des sujets très variés tels la santé, la famille, l'emploi du temps, les réseaux sociaux, etc. Les taux de réponse se maintiennent entre 80 % et 85 % et la méthode d'ITAO a été intégrée à l'enquête avec succès en 1993.

- L'enquête supplémentaire à l'EPA s'avère un choix judicieux lorsqu'il est essentiel d'effectuer des suivis sur place ou de conserver des liens avec les données de l'EPA à des fins d'analyse. Cependant, elle devient difficile à justifier lorsque l'interview téléphonique, qui est beaucoup moins coûteuse, répond aux besoins .

- Le coût des appels interurbains n'a cessé de diminuer au cours des cinq dernières années, ce qui rend les enquêtes téléphoniques très efficaces.

- Dans bien des cas, la durée des interviews et le type d'information à recueillir sont tels qu'une enquête supplémentaire risquerait de mettre en péril le succès de l'EPA.

- Comme nous l'avons mentionné précédemment, la couverture téléphonique est excellente au Canada et ne cesse de s'améliorer.

- La technique de CTA, en plus d'être efficace, se prête très bien à des procédures de sélection lorsque nous voulons qu'une enquête vise certaines personnes ayant des caractéristiques particulières à l'intérieur d'un ménage.

- Les enquêtes téléphoniques menées à partir du bureau central ou des bureaux régionaux offrent des environnements qui favorisent une gestion plus serrée du travail et de la qualité. Plus particulièrement, la méthode d'ITAO, utilisée dans

presque toutes les enquêtes téléphoniques depuis environ quatre ans, offre plusieurs avantages, y compris la vérification interactive, le cheminement automatique des questions, et surtout une méthode très efficace d'attribuer les appels aux intervieweurs et de gérer les rappels. Tous ces facteurs contribuent à réduire les coûts au minimum. De plus, le contrôle des interviews par une tierce personne a grandement contribué à l'amélioration de la qualité de l'interview et de l'instrument de collecte.

Le *tableau 4* présente une liste non exhaustive mais représentative des enquêtes ITAO effectuées depuis le début des années 90.

Tableau 4
Enquêtes spéciales ITAO

Enquête	Année	Base de Sondage	Taille de l'échantillon	Longueur moyenne de l'interview (minutes)	Nombre d'écrans
Décrocheurs	1991	Liste	15 000	35	750
Violence envers les femmes	1993	CTA	12 300	80	700
Enquête nationale sur les métiers d'apprentissage	1994	Liste	19 900	29	*
Enquête sur la population active du secteur culturel	1994	Liste	19 000	60	1 300
Enquête sur le tabagisme	1994	CTA	21 400	20	500
Enquête Canadienne sur l'alcool et les autres drogues	1995	CTA	35 000	36	*
Enquête sur l'exposition au soleil	1995	CTA	5 900	15	400
Enquête sur les personnes touchant des prestations d'invalidité du Régime de Pension du Canada	1995	Liste	6 700	30	600
Suivi sur les décrocheurs	1995	Liste	9 400	45	800
Suivi auprès des diplomés de 1990	1995	Liste	34 000	26	*
Enquête nationale sur l'utilisation des véhicules privés	1995/96	Liste	37 000	10	700
Enquête sur l'asthme	1996	Liste	4 000	30	600
Enquête nationale sur l'usage des médias électroniques	1996	CTA	4 000	30	900
Suivi sur l'enquête canadienne par panel sur l'interruption d'emploi	1996	Liste	10 000	40	1 000

* Non-disponible

Quelques tendances de fond se dégagent des leçons tirées de ces enquêtes.

- L'**infrastructure** nécessaire au soutien des enquêtes ITAO est beaucoup plus légère que celle requise pour les enquêtes- ménages sur le terrain. Elle s'adapte aussi bien à de très petites enquêtes spéciales (3 000 à 5 000 unités) qu'à des enquêtes téléphoniques de grande envergure (35 000 unités ou plus), et elle

contribue à l'efficacité de cette méthode. Par ailleurs, l'intégration de contrôles interactifs à la collecte ainsi que l'élimination des opérations de saisie et la réduction des vérifications et des suivis subséquents ont mené à une amélioration considérable de **l'efficience des opérations**.

- Dans le passé, nous hésitions beaucoup à confier à nos enquêteurs des interviews téléphoniques de plus de 10 à 15 minutes. Or, nous avons observé une évolution importante à cet égard. En effet, comme l'indique le tableau 4, il est courant de voir des **interviews d'une durée moyenne** de 30 à 45 minutes. Peut-être avions-nous sous-estimé le niveau de tolérance des répondants. La crédibilité dont jouit Statistique Canada, tout particulièrement depuis cinq ans, constitue également un facteur important.

- À titre d'exemple, la durée moyenne des interviews de l'Enquête sur la violence envers les femmes et de l'Enquête sur la population active du secteur culturel a été de l'ordre de 60 à 80 minutes. Ces expériences montrent clairement qu'il est possible de "pousser l'audace" dans les cas particuliers où les sujets traités sont d'un intérêt marqué pour les répondants.

- L'ITAO a contribué à faire reculer les limites du possible en ce qui a trait à la **complexité de la structure des questionnaires**. Le contrôle et le cheminement informatisés des questions ouvrent de nouveaux horizons. En effet, la structure des questionnaires informatisés intégrés à l'ITAO atteint des niveaux impossibles à réaliser avec des questionnaires traditionnels. Cette nouvelle réalité est illustrée au tableau 4 par le nombre d'écrans qui ont dû être programmés pour chacune des applications, soit plus de 500 dans bien des cas et même 1 300 dans un cas particulier.

- Nous avons aussi noté une nette progression au niveau de **la nature délicate des sujets** traités lors d'une enquête téléphonique. Une enquête sur la violence envers les femmes ou sur l'alcool et les drogues aurait été impensable dans le contexte des années 70 ou 80.

5. ENQUÊTES AUPRÈS DES AGRICULTEURS ET DES ENTREPRISES

5.1 Enquêtes agricoles

Jusqu'au début des années 80, la poste était la principale méthode d'envoi de questionnaires aux agriculteurs pour recueillir l'information nécessaire à la production des indicateurs statistiques. Chaque année, on mettait à la poste environ

200 000 questionnaires et les estimations étaient basées sur des taux de réponse de l'ordre d e 25 %. Vers les années 70, nous avons commencé à faire l'essai de méthodes probabilistes dans le secteur de l'agriculture, mais ce n'est que lors du remaniement des enquêtes, à la suite du recensement de 1981, que les méthodes non probabilistes ont graduellement été remplacées par des méthodes d'enquête probabilistes.

Parallèlement, étant donné que l'échantillonnage était plus ciblé et de taille plus restreinte, nous avons commencé à faire des suivis téléphoniques au niveau de la collecte, afin de préserver la représentativité de l'échantillonnage. Or, au début des années 90, l'avènement des méthodes d'interview assistée par ordinateur no us a incité à repenser notre stratégie de collecte.

Comme le montre le *tableau 5*, plusieurs enquêtes agricoles exigent la collecte de données auprès d'un assez grand nombre d'exploitations agricoles sur une très courte période de temps. Par ailleurs, les envois postaux avec suivi téléphonique étaient très mal adaptés à cette nouvelle stratégie d'enquête. La durée des interviews étant généralement très courte (de 8 à 12 minutes) et les délais de publication très serrés, il a été décidé d'adopter la méthode d'ITAO pour la grande majorité des enquêtes agricoles. Avec un certain recul et après trois ans d'expérience de l'ITAO, nous pouvons dire sans hésitation que cette décision a été la bonne. Très rapide et efficace, le processus d'enquête s'en trouve grandement simplifié et aussi amélioré par l'introduction de contrôles interactifs.

Tableau 5
Enquêtes agricoles ITAO
Quelques exemples pour l'année 1995

Enquête	Taille de l'échantillon	Longueur moyenne de l'interview (minutes)	Période de collecte (jours)	Nombre d'intervieweurs	Taux de réponses
Enquête sur les fruits et légumes	20 000	8	16	43	97%
Enquête de juin sur les fermes	28 000	9	10	99	90%
Enquête de novembre sur les fermes	27 000	10	15	73	93%
Enquête de juillet sur le bétail	27 000	8	12	81	97%
Enquête sur les serres et les pépinières	3 200	22	17	17	94%

5.2 Enquêtes auprès des entreprises

Pour la plupart des enquêtes auprès des entreprises, on a encore recours aux envois postaux comme méthode initiale de collecte, puisque cette méthode demeure la façon la plus économique de recueillir l'information statistique pour ce type d'unités. Bien sûr, cette méthode ne permet pas à elle seule d'obtenir des taux de réponse satisfaisants et elle exige des suivis téléphoniques. Or, comme nous l'avons mentionné précédemment (voir *section 2.3 et tableau 1*), le pourcentage d'unités déclarantes par téléphone est passé de 45 % à 63 % au cours des cinq dernières années.

Plusieurs raisons expliquent cette augmentation.

- Dans le passé, des cartes postales de suivi étaient envoyées afin d'inciter les répondants à retourner les questionnaires par la poste, mais cette méthode produisait des résultats très décevants et occasionnait des retards. Cette méthode a été complètement abandonnée au profit d'une stratégie de suivis téléphoniques beau coup plus rapides et agressifs. A l'occasion, nous effectuons un suivi par télécopieur entre l'envoi postal et le suivi téléphonique lorsque le calendrier le permet. Ces suivis sont entièrement informatisés et peuvent produire des résultats intéressants, particulièrement lorsque le retour s'effectue également par télécopieur.

- Plusieurs enquêtes mensuelles ont une importance capitale en ce qui a trait à la production des estimations du produit intérieur brut. Les taux de réponse ciblés de ces enquêtes sont généralement de l'ordre de 95 %, alors que temps alloué à la collecte est très court. Le succès d'une telle opération dépend donc largement d'un bon ordonnancement du travail et des ressources disponibles. Dans le cas de ces enquêtes, nous procédons à une analyse des tendances de réponse de chaque entreprise et nous envoyons un questionnaire par la poste seulement aux unités qui répondent fidèlement sans suivi téléphonique. Ainsi, les entreprises qui exigent un suivi téléphonique régulier sont tout simplement retirées de la liste des envois postaux et on procède immédiatement à la collecte par téléphone au début de la période d'enquête, ce qui permet d'équilibrer la charge de travail tout au long de la période de collecte. Par exemple, le taux d'interview téléphonique est d'environ 50 % pour l'Enquête mensuelle auprès des commerces de gros et de détail et de 90 % dans le cas de l'Enquête mensuelle auprès des manufactures.

- Comme nous l'avons fait pour les enquêtes agricoles, nous sommes en train d'intégrer l'ITAO aux enquêtes-entreprises, mais seulement pour les suivis téléphoniques. Les données des questionnaires retournés par la poste sont saisies de la façon traditionnelle. Cependant, les rejets aux contrôles effectués sur ces

questionnaires sont intégrés aux suivis téléphoniques assistés par ordinateur, ce qui permet de bénéficier des avantages de cette méthode.

Pour ce qui est des enquêtes spéciales auprès des entreprises, nous effectuons généralement des contacts préliminaires par téléphone afin de connaître le ou les représentants de l'entreprise les plus susceptibles de fournir l'information requise. Nous profitons également de l'occasion pour informer les contacts de la tenue de l'enquête et de l'importance d'y participer et de remplir le questionnaire qui leur parviendra par la poste.

6. CONCLUSION

Comme nous l'avons démontré, l'interview téléphonique occupe une place importante dans le programme d'enquêtes de Statistique Canada. Même si l'interview sur place demeure, selon nous, la façon la plus sûre d'établir le contact initial pour l'Enquête sur la population active, il n'en demeure pas moins que l'interview téléphonique est une méthode fiable et efficace pour établir un contact subséquent et, dans certains cas, le contact initial.

Pour ce qui est du secteur de l'agriculture, l'expérience a démontré que l'ITAO est une méthode très bien adaptée à ce type d'enquête. Quant aux enquêtes-entreprises, l'interview téléphonique est un outil essentiel à l'obtention d'un taux de réponse suffisant pour répondre aux besoins statistiques.

L'utilisation des méthodes d'interview téléphonique n'est pas sans soulever certaines questions ou inquiétudes. Par exemple, la prolifération de ces enquêtes, qu'elles soient pour le compte d'un organisme de statistique gouvernemental ou d'une firme de sondage, est-elle en train de saturer la population au point de miner la bonne volonté des répondants ? Notre crédibilité en tant qu'organisme de statistique national suffira-t-elle à maintenir des taux de réponse acceptables pour nos enquêtes téléphoniques dans le futur ? Voilà des questions qui influeront grandement sur l'avenir des enquêtes par téléphone.

À court et moyen termes, nous n'entrevoyons pas une diminution de l'usage des interviews téléphoniques. Les avantages qu'elles offrent sur le plan de la souplesse et de la rapidité d'exécution en font une approche de prédilection pour quiconque cherche à équilibrer les coûts, la qualité et les délais.

À plus long terme, la transmission électronique représente sûrement une voie d'avenir pour la collecte de données. Supplantera-t-elle complètement les interviews téléphoniques ? Cela est très peu probable. A notre avis, les organismes de statistique devront investir afin d'accroître au maximum la souplesse des modes de collecte pour encourager la participation.

BIBLIOGRAPHIE

CATLIN, G., INGRAM, S., « The effects of CATI on cost and data quality », *Telephone Survey Methodology*, éditions R. Groves et coll., 437-350. New York: Wiley, 1988.

CHOUDHRY, G.H., Results from telephone interviewing experiment in non-self-representing areas of the Labour Force Survey, Document interne, Statistique Canada, 1984.

DREW, J.D., « Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada », *Techniques d'enquêtes*, Vol. 17, n° 1, 63-75, Statistique Canada, 1991.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II, WAKSBERG, J. (éds), *Telephone Survey Methodology*, New York: Wiley, 1988.

LAFLAMME, F., Étude comparative entre trois différentes populations visées par l'EPA selon le type de service téléphonique, Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada, 1990.

MIURHEAD, R.C., GOWER, A.R., NEWTON, F.T., « The telephone experiment in the Canadian Labour Force Survey », *Survey Methodology*, 1, 158-180, 1975.

Session 1

Les questionnaires et réponses aux enquêtes

CONCEPTION ET ÉVALUATION DE QUESTIONNAIRES

France Bilocq

1. INTRODUCTION

Plusieurs organismes publics et privés effectuent chaque année des enquêtes auprès des personnes, des ménages et des entreprises. Ces enquêtes couvrent un ensemble de sujets variés et sont généralement réalisées soit par la poste, par entrevue téléphonique ou en face à face. Le questionnaire d'enquête est le principal véhicule de communication entre ces organismes et les enquêtés. La qualité des données d'une enquête est donc directement liée à la qualité du questionnaire utilisé pour la collecte. Un plan de sondage optimal et un système de traitement de données hautement efficace ne pourront fournir des données d'enquête de qualité à partir d'un questionnaire mal conçu. La conception et l'évaluation d'un questionnaire sont par conséquent des étapes très importantes dans le développement d'une enquête.

Le domaine d'étude de la conception et évaluation de questionnaires est très vaste et a fait l'objet de plusieurs publications. L'objectif de ce document n'est pas de décrire de façon exhaustive tous les points dont il faut tenir compte en conception et évaluation de questionnaires. Dans un premier temps, on verra l'importance de bien définir les objectifs d'une enquête pour faciliter la conception d'un questionnaire. Ensuite, un certain nombre de principes de base concernant la formulation de questions seront présentés en tenant compte des rôles spécifiques que peuvent jouer les enquêtés et les enquêteurs lors d'une collecte. Finalement quelques méthodes d'évaluation de questionnaires seront décrites. Il faut noter que les exemples utilisés dans le document concernent les enquêtes ménages. Toutefois, dans l'ensemble, les principes de bases décrits sont suffisamment généraux pour s'appliquer aussi aux enquêtes entreprises.

2. LES OBJECTIFS D'UNE ENQUÊTE

Avant de décider de faire une enquête il est de mise de consulter les travaux, enquêtes et études ayant porté sur le même sujet afin d'en tirer le maximum d'information et de profiter de l'expérience acquise (bonne ou mauvaise) lors de ces projets. Une fois le projet d'enquête décidé, il convient dans un premier temps

d'identifier les hypothèses à l'étude et de définir les objectifs de l'enquête en collaboration avec les commanditaires et les experts du sujet.

Les objectifs d'une enquête sont exprimés en termes de concepts et le questionnaire devient l'outil par lequel ces concepts sont mesurés. Le but à atteindre en concevant un questionnaire est de faire en sorte que les questions posées produisent une réponse fiable et qu'elles mesurent de façon adéquate ce que l'on veut étudier.

Pour ce faire, il faut décrire avec précision les concepts reliés aux objectifs. Supposons qu'on veuille faire une étude sur les conditions de logement des ménages à faible revenu. Cet objectif donne une idée générale de l'enquête mais nécessite plus de précision. Les concepts doivent être mieux définis. Par exemples : qu'entend-t-on par conditions de logement? (date de construction, nombre de pièces, nombre de personnes y vivant, coût du logement etc.), qu'est-ce qu'un revenu? (salaire, dettes, placements), sur quelle base mesure-t-on le revenu? (net/brut, hebdomadaire/mensuel/ annuel) à quel niveau? (individu/ménage), etc.

La réponse à ces questions dépend de l'utilisation finale des données. Il faut donc identifier les utilisateurs potentiels des données de même que l'utilisation et les analyses statistiques qui seront éventuellement effectuées à partir de ces données.

Selon Fowler « *Une des principales sources de mauvaises questions d'enquête est que la transition entre la précision d'un objectif et une question ne se fait pas. L'objectif est immédiatement écrit sous forme de questions* ».

Afin de faciliter la conception d'un questionnaire il est suggéré d'effectuer les étapes suivantes :

- Étudier les sources d'information existantes
- Écrire la liste des objectifs de l'enquête
- Établir le plan d'analyse

Une fois ces tâches accomplies on peut commencer à concevoir le questionnaire d'enquête en gardant en tête le principe suivant (Fowler) :

Si on ne peut relier une question à un objectif et à un rôle dans le plan d'analyse alors cette question doit être reformulée ou enlevée du questionnaire d'enquête.

Une fois le questionnaire conçu, l'enquêté et l'enquêteur deviennent les principaux acteurs du scénario de collecte d'information. On verra dans les sections suivantes comment, lors de la conception d'un questionnaire on peut minimiser les erreurs de mesure imputables à ces acteurs.

3. L'ENQUÊTÉ ET LE PROCESSUS COGNITIF DE RÉPONSE À UNE QUESTION

En répondant à une question, l'enquêté effectue une série de tâches que l'on appelle le processus cognitif de réponse à une question. Ce processus cognitif se décompose en cinq étapes principales (Sudman et al. 1996). Un diagramme représentant le processus cognitif se trouve en annexe.

Étape 1 : Interprétation

Dans un premier temps, l'enquêté décode la question. Il en interprète le sens. Cette tâche implique deux opérations qui sont reliées entre elles : la première, comprendre les mots utilisés dans la question et la deuxième, donner un sens à la question posée afin de fournir une réponse adéquate.

Étape 2 : Extraction, recherche de l'information

Une fois l'étape d'interprétation terminée, l'enquêté fait appel à sa mémoire pour extraire l'information nécessaire. Soit l'information est disponible immédiatement (par exemples : nom, âge, sexe, etc.) soit il doit reconstruire les événements, souvenirs ou connaissances reliés à la question. Dans certains cas l'enquêté devra consulter des documents pour obtenir la réponse.

Étape 3 : Choix et format de la réponse

Ensuite l'enquêté choisit parmi les réponses possibles celle qu'il considère adéquate. Il prend une décision quant à la réponse à donner à la question. Par exemple, il choisit l'événement, le souvenir ou les connaissances qu'il résumera pour répondre à la question. Dans le cas d'une question fermée, il choisit la catégorie qui correspond le mieux à sa réponse.

Étape 4 : Modification

En présence d'une question qu'il juge "sensible ou délicate" l'enquêté peut parfois modifier sa réponse afin de bien paraître ou pour se classifier dans les « normes sociales ».

Étape 5 : Communication

À la dernière étape, l'enquêté communique par écrit ou oralement la réponse qu'il a choisie.

Afin de faciliter la compréhension du processus cognitif de réponse, les étapes ont été présentées en ordre séquentiel. Toutefois, cet ordre ne représente pas nécessairement celui dans lequel l'enquêté les effectue.

Comme l'enquêté est la personne la plus importante dans une enquête, il est essentiel lors de la conception d'un questionnaire, de tenir compte du processus cognitif de réponse des enquêtés. En facilitant leurs tâches cognitives, on maximise les chances d'obtenir des réponses adéquates aux questions posées. Quatre des cinq étapes du processus cognitif seront, dans les paragraphes suivants, présentés sous forme de défis à relever lors de la conception d'un questionnaire.

DÉFI 1 : Lié à l'étape interprétation (vocabulaire, concepts et sens de la question)

Une question d'enquête doit être construite pour que le vocabulaire utilisé soit compris et interprété de la même façon par les enquêtés. De plus, le sens donné à la question devrait être le même que celui voulu par le concepteur. Par exemple (Fowler), on a demandé à des enquêtés, combien de jours durant la semaine précédent l'enquête, avaient-ils mangé beurre? Plusieurs personnes utilisent le terme beurre pour désigner de la margarine. À ce titre les enquêtés ont parfois inclus ou parfois exclu la consommation de margarine dans leur réponse. Lorsque la question a été posée à nouveau mais en excluant explicitement la margarine, le nombre de jours de consommation de beurre a diminué de 20% par rapport à l'autre question moins explicite. On se rend compte que malgré l'utilisation de mots simples, une question peut avoir plusieurs sens.

L'utilisation de mots difficiles devrait être compensée par la présence d'une définition dans le questionnaire ou dans le guide d'instruction. De même les concepts compliqués devraient faire l'objet de questions multiples. La tâche du répondant s'en trouvera ainsi simplifiée. Par exemple, dans le questionnaire ménage du Panel Européen français, la question concernant les placements financiers se décompose en quatre sous-questions et ce afin que l'enquêté puisse distinguer chaque type de placements.

Pour éviter toute confusion dans l'interprétation d'une question et de la réponse à donner il faut éviter l'emploi d'une négation ou d'une double négation.

Exemple : N'êtes vous pas contre le fait de disposer de poubelles spéciales permettant de recycler le verre dans votre quartier?

DÉFI 2 : Lié à l'étape extraction, recherche de l'information pertinente

Une fois la question conçue afin que l'enquêté comprenne bien ce qu'on lui demande, il faut se demander si l'enquêté connaît la réponse et s'il s'en souvient. À moins que l'objectif de la question soit de mesurer des connaissances/compétences,

le fait qu'un répondant ne connaisse pas la réponse à une question constitue une source d'erreur. L'enquêté se retrouve face à diverses situations :

i. L'effet connaissance

On fait souvent l'hypothèse en conception de questionnaire que l'enquêté connaît les réponses à toutes les questions. Il y a des sujets plus difficiles que d'autres et un enquêté peut ne pas pouvoir y répondre par manque de connaissance.

On peut faire l'hypothèse qu'un enquêté peut en général rapporter des faits le concernant : visites médicales, alimentation, budget, etc. Toutefois, il rapportera avec beaucoup moins de précision des informations détenues par d'autres personnes. Par exemple, un enquêté pourra difficilement faire la description des tests médicaux que lui a fait subir son médecin durant les six derniers mois. Il est donc important de poser à l'enquêté des questions pour lesquels il détient ou a accès à l'information demandée. Dans le même ordre d'idée, lorsqu'un enquêté répond pour un autre (proxy), il peut ne pas connaître les réponses à toutes les questions.

Parfois l'enquêté détient l'information reliée à la question mais pas sous la forme demandée. Il existe de bons exemples dans le domaine financier : un enquêté sait qu'il a des placements financiers mais peut ignorer le terme exact les désignant. Si le terme exact est utilisé dans la question alors l'enquêté peut ne pas connaître la réponse.

Il faut donc tenir compte du niveau de connaissance lié à une question pour que l'enquêté se retrouve dans une situation où il pourra fournir une réponse adéquate aux questions posées.

ii. L'effet mémoire.

Il est plus facile pour un enquêté de se souvenir d'un événement récent, d'un événement important ou d'un événement en accord avec sa façon de penser. Pourtant on lui demande souvent de rapporter des événements qui ne font partie d'aucune de ces catégories. En utilisant une période de référence courte on diminue la possibilité d'erreur de la part de l'enquêté. De plus, en posant des questions multiples ou en proposant des associations d'idées on stimule la mémoire de l'enquêté.

iii. L'effet télescopage.

Le télescopage se produit lorsque l'enquêté se souvient d'un événement mais a de la difficulté à le situer dans le temps. Il le rapporte alors dans une période qui n'est pas la bonne. Voici quelques solutions qui peuvent minimiser l'effet de télescopage :

- Utiliser un calendrier

- Utiliser des repères : anniversaire, mariage, décès ou autre événement important

- Utiliser un carnet ou un journal dans lequel l'enquêté note les événements

- (exemples: emploi du temps, alimentation, dépenses, transports)

- Rencontrer plusieurs fois l'enquêté, par exemple :

Au 1er contact, on demande à l'enquêté de rapporter ses visites médicales qui ont eu lieu depuis deux mois.

On lui spécifie aussi que lors de la prochaine entrevue (dans six mois), on lui demandera de rapporter les visites médicales qu'il effectuera durant les six prochains mois.

Il y a trois avantages à cette méthode. Le premier, la période de temps et les événements à rapporter sont clairement définis. Ensuite les visites médicales rapportées lors du premier contact permettront de compenser l'effet de télescopage du deuxième contact. Et finalement, le premier contact rend l'enquêté plus attentif et donc meilleur répondant lors du deuxième contact.

Chacune des suggestions précédentes possède des avantages et des inconvénients. Par exemple, présenter un calendrier aux enquêtés rallonge le temps d'entrevue et avertir l'enquêté qu'on veut de l'information sur un certain type d'événement peut modifier son comportement. Il faut juger selon les situations et choisir la solution qui permettra une meilleure collecte d'information.

DÉFI 3 : Lié à l'étape choix et format de la réponse

Pour concevoir des questions auxquelles les enquêtés peuvent répondre il faut respecter quelques règles.

- Il est important de bien définir le contexte d'une question, de fournir des points de référence (période, lieu, unité de mesure etc.) afin d'indiquer à l'enquêté la forme qu'est supposée prendre la réponse.

Exemple :

Question :	Depuis quand habitez-vous la ville de Lyon?
Réponses possibles :	Depuis 1948 Depuis l'âge de 10 ans Depuis toujours
Meilleure question :	Depuis combien d'année habitez-vous la ville de Lyon?

La dernière question spécifie l'unité de mesure et facilite la tâche de l'enquêté qui fournira une réponse plus rapide.

- L'unité de mesure de la réponse doit correspondre au type d'information que peut fournir un enquêté sinon il risque d'y avoir erreur de mesure. Par exemple, si on demande à combien de km se situe l'hôpital le plus proche de votre logement, l'enquêté peut savoir où est l'hôpital mais ne connaît pas la distance exacte en km. En région urbaine, il est parfois plus approprié de demander le temps et le mode de transport que la distance parcourue.

- Chaque question doit contenir une seule question et une idée principale sinon l'enquêté sera confus et ne saura pas à laquelle des questions il doit répondre. Il sera de plus difficile de tirer des conclusions concernant la réponse fournie.

  Exemple : Votre entreprise prévoit-elle offrir à ses employés un plan d'assurance dentaire et payera-t-elle100% des primes?

- Dans le cadre d'une question fermée, il faut fournir à l'enquêté un choix de réponse clair, exhaustif et sans chevauchement que les réponses possibles soient numériques ou non.

DÉFI 4 : Lier à l'étape modification (sujets sensibles et normes sociales)

Dans le cas de sujets jugés sensibles comme la consommation de drogue, d'alcool, l'avortement, la religion, les comportements sexuels, etc. des études ont démontré qu'un enquêté aura tendance à ne pas répondre ou à modifier sa réponse pour cacher la vérité. La notion de sensibilité n'est cependant pas la même pour chaque individu. De manière générale, on peut dire qu'un sujet devient sensible pour une personne lorsqu'elle est concernée par le sujet. Par exemple, en milieu défavorisé, poser une question sur la présence ou non d'un système de chauffage électrique dans un logement sans même se préoccuper de savoir si les occupants ont l'électricité peut devenir pour un enquêté une question sensible (Dechaud-Rayssiguier). L'enquêté à qui ont pose une question sensible peut se sentir jugé et avoir peur des répercussions que peut lui apporter sa réponse. En posant de telles questions sans prendre de précaution on court le risque d'irriter l'enquêté, d'obtenir une non-réponse totale ou partielle, d'obtenir une réponse fausse et de créer un souci de cohérence artificiel chez l'enquêté tout au long du questionnaire.

L'intégration de procédures de collecte différentes et une formulation prudente des questions sont deux éléments essentiels pour recueillir de bonnes données sur les sujets sensibles. Voici une liste de stratégies pouvant être utilisées pour favoriser une réponse précise aux questions portant sur des sujets sensibles.

- Rendre l'enquêté à l'aise de fournir une réponse « différente »

- Justifier les questions
- Assurer la confidentialité
- Utiliser des intervalles comme choix de réponses
- Dans le cas d'une entrevue, réduire le rôle des enquêteurs dans le processus de collecte en ayant une partie du questionnaire (anonyme) rempli confidentiellement par l'enquêté.

  Par exemples : - Donner à l'enquêté un questionnaire papier à compléter et à remettre sous enveloppe scellée à l'enquêteur.

  - Faire remplir une section du système informatique de collecte par l'enquêté lui-même.

  - Utiliser un message vocal avec téléphone à touche.

- Utiliser des techniques de réponse anonymisée

Finalement, Il faut adopter une approche neutre et formuler une question de manière à ne pas suggérer la réponse à l'enquêté. Sinon, l'enquête sera biaisée et on ne mesurera pas l'opinion des répondants correctement.

Exemple : Les études scientifiques montrent que l'usage du tabac est à l'origine de coûts sociaux importants. Êtes-vous en faveur d'une législation interdisant de fumer dans les lieux publics?

Dans les paragraphes précédents, tous les défis à relever qui tiennent compte du processus cognitif n'ont pas été énumérés. Ceux qui sont décrits donnent toutefois une bonne idée des principes de base à appliquer en conception de questionnaires qu'il s'agisse d'une enquête ménage ou d'une enquête entreprise.

4. L'ENQUÊTEUR

Dans les enquêtes par entrevue (téléphone et face à face), les enquêteurs sont sur la ligne de front. Ils assurent la liaison entre les organismes effectuant les enquêtes et les enquêtés. Ils ont une grande responsabilité et il est parfois possible de faciliter leurs tâches au moment de la conception d'un questionnaire.

Les tâches d'un enquêteur sont nombreuses et complexes. Par exemple, dans le cadre des enquêtes ménage menées par entrevue en face à face, l'enquêteur doit après avoir effectué les tâches de repérage du logement, contacter l'enquêté, lui décrire l'enquête et ses objectifs de manière à susciter sa participation, lire les questions, écouter la réponse, déterminer si la réponse est adéquate, faire au besoin

un suivi, répondre aux questions éventuelles de l'enquêté et saisir la réponse. Lors d'une entrevue, l'enquêteur devient tour à tour émetteur et capteur d'un message. Il doit pouvoir transmettre correctement son message et être à l'écoute de l'enquêté pour saisir la réponse fournie. Lors de la conception du questionnaire, la lecture des questions, des instructions, les sauts de question(s) (aiguillage) et la saisie doivent être simplifiés au maximum afin que l'enquêteur se concentre sur la partie la plus difficile de son travail : l'échange d'information avec l'enquêté.

Voici quelques exemples :

- Identifier clairement et uniformément les instructions pour l'enquêteur et les phrases qu'il doit lire à l'enquêté (majuscules, zone ombrée, touche d'appel etc.).

- Utiliser toujours la même convention pour indiquer l'aiguillage (automatique dans le cas d'un système informatisé de collecte)

- Lorsqu'un choix de mots s'impose dans une question faire en sorte que les choix soient identifiés de manière uniforme dans le questionnaire par exemples : il/elle, conjoint/conjointe etc.). L'objectif étant d'avertir l'enquêteur qu'il doit faire un choix. Ce choix devrait être automatique dans le cas d'une collecte assistée par ordinateur.

NOTE: Les deux premiers points s'appliquent aussi dans le cas d'une collecte par la poste.

Il y a d'autres aspects en conception de questionnaire (formulation de question, sujets sensibles etc.) dont il faut tenir compte lors de l'élaboration d'un questionnaire prévu pour une collecte effectuée par enquêteurs. Ces derniers ont une expérience "du terrain" que peu de personnes possèdent et dont il faut tenir compte lorsque c'est possible de le faire. On verra dans la section 5.4 comment mettre à profit cette expérience. D'un autre côté il existe des contraintes parfois incontournables en conception de questionnaires. Il s'agit alors d'établir une collaboration de travail entre les personnes concevant le questionnaire et celles qui les utilisent c.-à-d. les enquêteurs. Les contraintes de travail de chacun seront mieux comprises de part et d'autre et on pourra ainsi maximiser la qualité d'un questionnaire d'enquête.

5. ÉVALUATION DE QUESTIONNAIRES

On a vu dans les sections précédentes l'importance de tenir compte des tâches de l'enquêté et de l'enquêteur au moment de l'élaboration d'un questionnaire. Il est tout aussi important de développer des outils permettant d'évaluer un questionnaire.

C'est à dire de vérifier que le vocabulaire est simple, que les questions ne sont pas trop difficiles à interpréter et à répondre, que la tâche demandée est réalisable etc. Différentes techniques sont utilisées pour évaluer un questionnaire. Dans les prochains paragraphes les techniques suivantes seront décrites : le groupe de discussion, l'entrevue en profondeur, le pré-test et les séances de discussion avec les enquêteurs. Ces différentes techniques sont présentées selon l'ordre chronologique dans lequel elles devraient survenir dans une enquête. Plusieurs autres techniques existent dont celles relevant de la psychologie cognitive. Elles sont documentées dans la littérature.

5.1 Groupe de discussion (Focus group)

Les groupes de discussion sont utilisés depuis longtemps par plusieurs organismes statistiques pour l'évaluation de questionnaires. C'est une technique d'entrevue qui consiste à réunir des personnes (participants) pour discuter par exemple d'un problème commun, d'une idée ou d'un concept.

L'efficacité d'un groupe de discussion est basée sur l'hypothèse que la discussion entre les participants génère des échanges qui n'auraient pu survenir lors d'une entrevue entre deux personnes. Les groupes de discussion avec les enquêtés sont utiles en évaluation de questionnaires pour entre autres :

- Étudier les hypothèses concernant les sujets couverts par l'enquête
- Explorer les connaissances, opinions, attitudes, perception etc. des participants envers ces sujets
- Évaluer la compréhension du vocabulaire et des concepts utilisés dans le questionnaire
- Déterminer si les « tâches » (procesus cognitif) à accomplir pour répondre à une question sont réalisables (mémoire, calcul, ordonnancement, etc.)
- Évaluer la confiance des participants concernant les politiques de confidentialité et l'utilisation des données.

Les résultats d'un groupe de discussion ne peuvent être projetés sur l'ensemble de la population. Toutefois, ils permettent d'identifier des problèmes que des enquêtés pourraient rencontrer avec le questionnaire. Le groupe peut aussi apporter des solutions à ces problèmes et ainsi aider à améliorer le questionnaire.

Le groupe de discussion est dirigé par un animateur dont les principales tâches sont de rendre à l'aise les participants, de favoriser les échanges, de donner la chance à

tous de s'exprimer et de choisir entre laisser la discussion suivre son cours et garder la conversation centrée sur les sujets d'intérêt. La composition du groupe influence les résultats qu'on peut en tirer. Selon le sujet à l'étude, on favorisera des groupes homogènes (sexe, niveau d'éducation, catégorie socio-professionnelle etc.) ou hétérogènes. On inclura des personnes ayant des expériences et des opinions diverses sur les sujets discutés. Dans la littérature on suggère de former des groupes de 8 à 10 participants. Lorsqu'il n'y a pas suffisamment de participants la dynamique de groupe n'est pas efficace et à l'inverse, le fait d'avoir trop de participants rend plus difficile la tâche de l'animateur.

Il est important, avant de mettre en place un groupe de discussion pour l'évaluation d'un questionnaire, d'identifier les points spécifiques à l'étude, par exemple le vocabulaire, la liste des réponses possibles, les effets de mémoire, le niveau de connaissances requis, etc. L'animateur demande alors aux participants de discuter principalement de leur expérience et de leur perception vis à vis des points à l'étude et aussi de tout autre point pouvant être soulevé dans la discussion.

Il importe de garder une trace de ce qui a été dit au cours de la discussion afin d'analyser les résultats. Une solution possible est d'avoir des observateurs qui prennent des notes. La meilleure stratégie est d'enregistrer (audio ou idéalement vidéo) le groupe. De cette façon, toutes les personnes impliquées dans un projet peuvent écouter/regarder l'enregistrement ensemble, discuter des résultats et au besoin écouter/visionner à nouveau des passages jugés importants.

Cette technique de discussion est une façon très efficace pour obtenir un grand nombre d'informations pertinentes pour une enquête. Le niveau d'investissement est minime et les résultats sont très utiles. Il serait donc à propos, que chaque enquête utilise cette technique en faisant appel à des enquêtés. Les groupes de discussion ne sont cependant que le début du processus d'évaluation d'un questionnaire. D'autres méthodes d'évaluation permettent d'examiner des aspects d'un questionnaire ne pouvant être abordés par un groupe de discussion.

5.2 L'entrevue en profondeur

L'entrevue en profondeur permet d'étudier plus en détail chacune des étapes du processus cognitif de réponse. La technique est la suivante : un observateur pose les questions du questionnaire à un participant, le laisse répondre et lui pose ensuite une série de questions sur les tâches qu'il vient d'accomplir c.-à-d. sur le processus cognitif de réponse. Si le questionnaire est auto-administré l'observateur prend en note la façon dont le participant rempli le questionnaire (les parties lues, la séquence de questions suivie, les hésitations etc.) et ensuite lui pose des questions. Voici quelques exemples de questions pouvant être posées aux participants :

- Paraphraser la question, les instructions
- Définir les mots, les concepts
- Demander le niveau de confiance dans leur réponse
- Expliquer les hésitations,
- Si la réponse demandait un calcul numérique, lui demander comment il a calculé sa réponse
- Si la réponse comportait un classement de modalités demander au participant de décrire le processus de décision de la réponse

Cette technique peut être effectuée de deux manières. On pose des questions immédiatement après chacune des questions du questionnaire ou on attend la fin de l'entrevue. La première a pour avantage qu'il est plus facile pour le participant de se rappeler de son processus de pensée tout de suite après avoir répondu à une question. Toutefois, on brise ainsi le rythme de l'entrevue de même que les liens qui peuvent exister entre les questions.

Tout comme les groupes de discussion cette technique sera plus productive si les points susceptibles de poser des problèmes aux participants ont été identifiés au préalable. Ensuite durant l'entrevue en profondeur, en même temps que les questions plus générales on pose les questions spécifiques concernant les points problèmes identifiés à l'avance.

Bien que les objectifs et les résultats des entrevues en profondeur et des groupes de discussion semblent similaires, les deux techniques sont complémentaires. Dans les deux cas on tente d'identifier les problèmes qui affecteront la compréhension des questions et la capacité des participants à y répondre de façon juste et consistante. Le groupe de discussion est une technique ou les problèmes sont abordés de manière générale. L'entrevue en profondeur est plus spécifique car elle se concentre plus spécialement sur les étapes du processus cognitif de réponse.

Il y a certaines difficultés que ni l'une ni l'autre de ces techniques ne peuvent identifier. Aucune ne teste les difficultés que pourrait rencontrer un enquêteur vis à vis du questionnaire et des questions (ordre des questions, aiguillage, instructions, espaces de réponse, vocabulaire, concept, lecture, compréhension etc). Le questionnaire doit par conséquent être testé par un échantillon d'enquêteurs. Finalement, il est essentiel de tester le questionnaire dans des conditions réelles d'enquête puisqu'il existe des situations ne pouvant surgir que lors de la simulation du processus de collecte réel.

5.3 Pré-test

Le pré-test consiste à effectuer une mini collecte auprès de quelques enquêtés dans des conditions réelles d'enquête. Si la collecte se fait par entrevue l'enquêteur signalera la présence de problèmes dans le questionnaire. Si la collecte se fait par la poste il faut informer les enquêtés qu'ils participent à un pré-test et leur fournir la possibilité de donner des commentaires pertinents sur le questionnaire. Il est important de saisir les données d'un pré-test afin de contrôler la cohérence entre certaines questions. Un pré-test identifie les problèmes mais ne permet pas d'en déterminer la cause ni les solutions éventuelles.

5.4 Séances de discussion avec les enquêteurs

Lors d'une collecte par entrevue, les enquêteurs sont en contact direct avec l'enquêté. Ils ont, en raison de leur expérience sur le terrain, un bon aperçu des bons et mauvais côté d'un questionnaire, des problèmes de passation d'un questionnaire et de la façon dont ce dernier influence la coopération des enquêtés. Leur collaboration est essentielle lors de la conception d'un questionnaire.

Des discussions avec les enquêteurs doivent se tenir au moment de la conception d'un questionnaire et après un pré-test. Elles permettent de déterminer les zones problèmes du questionnaire qui devront ensuite être améliorées. Fowler suggère de créer des questionnaires d'évaluation de questions destinés à être remplis par les enquêteurs à la suite d'un pré-test. Les enquêteurs y indiquent les problèmes rencontrés pour chaque question en suivant un schéma pré-établi. Ces questionnaires d'évaluation permettent aux concepteurs de rassembler des informations quantitatives concernant les problèmes soulevés et ainsi d'identifier les plus fréquents.

À la fin d'une enquête, un bilan général de collecte devrait être produit par les enquêteurs et faire partie intégrante de la documentation relative à cette enquête. Ainsi les bons et les mauvais côtés d'une enquête seront documentés et surtout pourront être pris en compte lors d'une prochaine enquête.

6. CONCLUSION

Ce document décrit un certain nombre de principes de bases pouvant être mis de l'avant au moment de la conception d'un questionnaire. Ces principes ont pour objet de faciliter les tâches des enquêteurs et des enquêtés et de minimiser les erreurs de mesure. Afin d'identifier les problèmes reliés à un questionnaire et éventuellement les corriger, il faut procéder à l'évaluation de ses questions et du processus cognitif

qu'elles engendrent et ce avant la collecte des données proprement dite. On a vu que certaines techniques d'évaluation nécessitent un minimum d'investissement en temps et argent et permettent d'améliorer l'efficacité d'un questionnaire. Finalement, l'expérience des enquêteurs doit être mise à profit tout au long de la conception et de l'évaluation d'un questionnaire de même qu'à la fin de la collecte. Il est donc essentiel que dans la planification d'une enquête, le temps et les ressources nécessaires soient alloués pour effectuer l'ensemble de ces activités. La qualité d'une enquête dépend de la qualité des informations recueillies du questionnaire d'enquête et de ses questions.

BIBLIOGRAPHIE

Forsyth, B. Lessler J.T., « Cognitive Laboratory Methods: A Taxonomy », *Measurement Errors in Surveys*, Chapt 20, Biemer P. et al., Wiley 1991

Dechaud-Rayssiguier, D., *Étude de cas, Enquête Extension de quartier, conditions de vie, situations défavorisées*, Document interne, Insee- DSDS, UMS, juin 1996

Fowler Floyd J. Jr., *Improving Survey Questions - Design and Evaluation*, Applied Social Research Methods Series, Volume 38. Sage Publications, Thousand Oaks, 1995

Groves, R.M., *Survey Errors and Survey Costs*, John Wiley & Sons, New York, 1989

Morgan David L., (ed.), *Sucessful Focus Groups - Advancing the State of the Art*, Sage Publications, Newbury Park, 1993

Nargundkar M.S., Platek R., *Qualitative Methods in Questionnaire Design*, I.S.I, 47th Session Paris, August 1989

Platek R., *Some Important Issues in Questionnaire Design Development*, JOS, Vol. 1, N° 2, 1985, pp. 119 -136

Schaeffer, Nora-Cate, *A decade of questions* , JOS 95:1

Sudman S., Bradburn N., Schwarz N., Thinking About Answers, *The application of cognitive processes to Survey Methodology*, Jossey-Bass Publishers, San Francisco, 1996

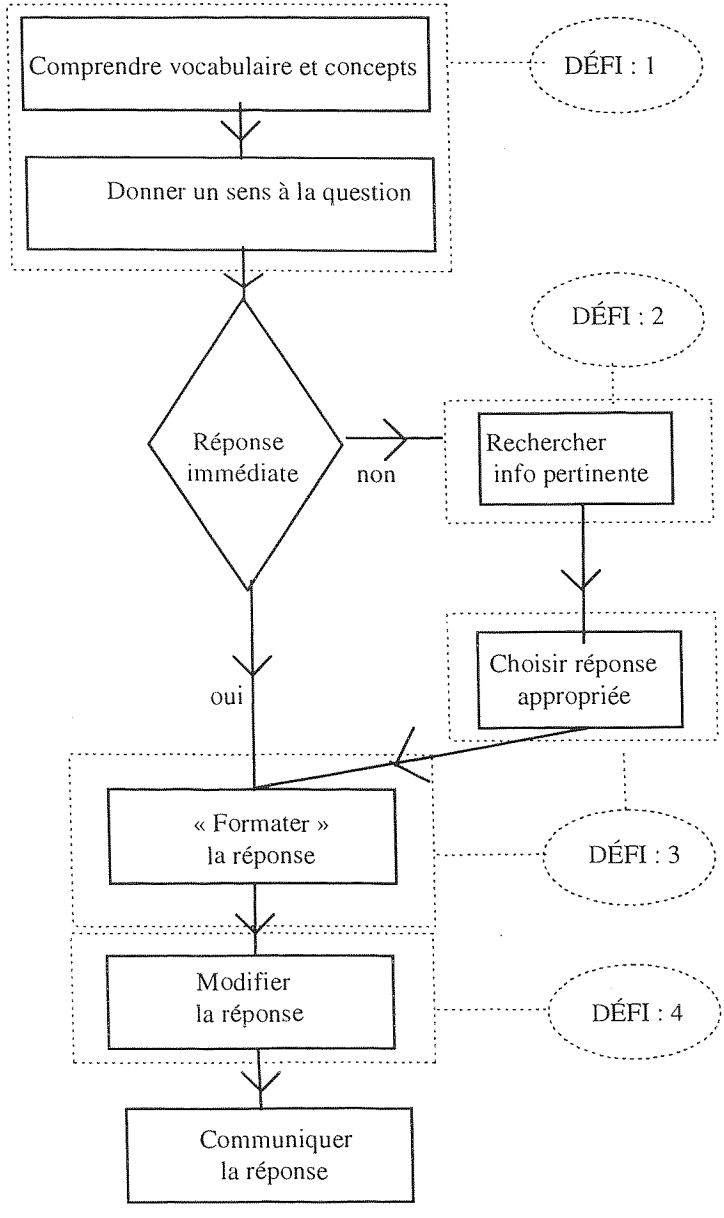
Willis G.B., *Cognitive Interviewing and Questionnaire Design: A Training Manual*, NHCS, Working Paper, Series, N° 7, 1994

Notes de cours, *Conception de questionnaires*, Statistique canada

Notes de cours, *Measurement in Surveys*, Cours TES, Eurostat, Statistics Sweden, Septembre 1996

ANNEXE

Représentation graphique du processus cognitif de réponse à une question d'enquête tirée de Sudman et al.1996, (traduction libre).



L'ÉLABORATION DES QUESTIONNAIRES DU 33ème RECENSEMENT DE LA POPULATION

Jacqueline Lacroix

I - La consultation pour le 33ème recensement

Bien que le contenu des questionnaires ait en apparence relativement peu changé au cours des 25 dernières années, des modifications interviennent à chaque recensement de manière à mieux prendre en compte les souhaits des principaux utilisateurs de données. Ces modifications se traduisent par l'introduction de nouvelles questions ou des changements de formulation.

Le processus de consultation pour le 33ème recensement a été entrepris au 4ème trimestre 1993, date qui peut être jugée trop précoce, mais qui se justifiait à l'époque par la perspective de réaliser ce recensement en 1997.

Administrations, en particulier services statistiques des ministères, organismes d'études et de recherches, aménageurs locaux, universitaires ... ont été consultés par le Département de la Démographie et les Directions Régionales de l'INSEE. Une réunion interformations du CNIS en décembre 1993 a permis de faire une synthèse des demandes formulées et d'établir un débat entre les utilisateurs.

Objectifs et limites de la consultation

Le recensement de 1990 ayant été considéré comme un succès, il a été décidé que le recensement de 1997 s'en inspirerait fortement. Aucune modification fondamentale dans l'organisation du recensement n'avait en effet été envisagée à l'issue des expérimentations menées dans le cadre de la préparation du RP 90. De plus, la proximité dans le temps des 2 recensements, telle qu'elle était prévue à ce moment, ne permettait pas de tester des innovations majeures.

On souhaitait de ce fait apporter seulement un nombre limité de modifications aux bulletins du recensement (bulletin individuel et feuille de logement). La consultation ne portait donc pas sur l'ensemble des questionnaires mais sur des thèmes précis où il semblait que le futur recensement pouvait apporter des éclairages nouveaux par rapport aux recensements antérieurs. En préambule, il était rappelé les conditions de réalisation du recensement avec leur impact sur le contenu des questionnaires.

Une technique de collecte qui impose un questionnement court

Le recensement de la population est une opération massive qui touche l'ensemble de la population. La méthode de collecte utilisée en France (dépôt-retrait) interdit d'utiliser des questionnaires lourds ou compliqués. Afin d'emporter l'adhésion de toute la population, il faut poser des questions simples - utilisant des concepts compréhensibles par tous - et considérées comme suffisamment banales pour ne pas risquer de refus de la part des recensés. Les renseignements qui peuvent être considérés comme trop personnels (vie privée) ou embarrassants (revenus) sont à éviter.

Côté présentation, le bulletin individuel du recensement est un recto-verso (2 pages) ce qui limite évidemment le questionnement. Un essai avait été fait en 1986 afin d'utiliser un bulletin individuel (BI) de 4 pages. Les problèmes de stockage, de difficultés de comptage manuel et la charge supplémentaire imposée à la population ont conduit l'INSEE à abandonner ce projet. Il n'était donc pas envisagé d'utiliser un BI de 4 pages en 1997.

De même, il n'était pas envisagé de modification importante dans l'organisation de la collecte comme la différenciation des questionnaires selon les populations (partie variable des questionnaires permettant un enrichissement des thèmes abordés par exploitation d'un sous-échantillon de taille significative). D'autres pays comme les Etats-Unis ou le Canada utilisent deux types de questionnaires (un léger et un lourd). En France, ne disposant pas de base de sondage des logements et ne faisant pas confiance a priori aux agents recenseurs pour opérer une sélection, on fait remplir à toutes les personnes le même bulletin individuel, alors même que les questions concernant l'activité professionnelle ne sont exploitées qu'une fois sur quatre. Le sondage se fait donc a posteriori au moment de l'exploitation.

Un échantillon, toutefois, est constitué a priori pour la réalisation de l'enquête Famille, principale source d'information sur l'évolution des structures familiales en France. Depuis 1954, cette enquête est associée au recensement dans un district sur 50. Jusqu'à présent, seules les femmes étaient concernées par l'enquête mais il est prévu de l'étendre aux hommes lors du prochain recensement de la population.

Un triple souci dans le choix des questions

Le contenu du questionnaire, proprement dit, doit satisfaire à trois objectifs parfois contradictoires : être adapté à la situation de l'époque, assurer la comparabilité d'un recensement à l'autre et être conforme aux recommandations internationales.

- **L'adaptation** à l'époque et aux besoins des utilisateurs a conduit à supprimer la question sur l'électricité dans le logement depuis 1962, à supprimer en 1990 les

questions sur le raccordement à un réseau d'eau, à ajouter en 1990, par exemple, des questions sur les emplois précaires...

- **La comparabilité** d'un recensement à l'autre conduit à maintenir le maximum de questions sans changement si elles sont toujours d'actualité.

- **Les recommandations de l'ONU** pour la campagne de recensement de 1990, répartissent les thèmes traités dans les recensements en 2 parties : les caractéristiques fondamentales et les principales «caractéristiques supplémentaires». Les recommandations de la campagne de 2000 ne sont pas encore définies. Si elles diffèrent de celles de 1990, elles ne pourront pas être prises en compte.

Les pays doivent tous poser les questions relatives aux caractéristiques fondamentales. Eurostat s'aligne sur les recommandations de l'ONU mais définit une liste de variables obligatoires pour les pays de l'union européenne : le programme commun de tableaux pour la campagne 1990 figurant dans sa directive reprend la liste des «caractéristiques fondamentales» et des principales «caractéristiques supplémentaires». Ces deux listes figurent dans l'annexe 1.

Le recensement français de 1990 comportait tous les sujets inclus dans les caractéristiques fondamentales sauf le régime de propriété (statut juridique du propriétaire), la présence de cuisine et le système d'adduction d'eau. Par contre, parmi les 48 «caractéristiques supplémentaires», la France n'en avait pas retenu 18 : la durée de résidence en France, l'année d'immigration, l'ethnie, la langue, la religion, le nombre d'enfants nés vivants, la date du mariage de la femme, le revenu, le lieu de l'école et le trajet induit, l'aptitude à lire et écrire, le montant du loyer, la disponibilité dans le logement d'eau chaude, d'électricité, de téléphone, la surface du logement, le type d'évacuation des eaux usées, l'isolation, les matériaux de construction.

De toute manière, la concision des documents de collecte ne permet pas de satisfaire à toutes les recommandations de l'ONU. Sans rentrer dans le détail des thèmes non abordés, on peut dire qu'une partie d'entre eux n'a pas été retenue car l'information existait déjà localement à partir de fichiers d'abonnés. C'est le cas pour la connexion à différents réseaux (eau, électricité, téléphone). Une autre partie concerne des thèmes «sensibles» que l'on préfère ne pas aborder en France (religion, ethnie) ou seulement par le biais d'enquêtes (revenus). Enfin, les questions se rapportant à la vie familiale (mariage, enfants) sont traitées de façon détaillée dans l'enquête Famille, ce qui est compatible avec les recommandations de l'ONU.

La prise en compte de la nomenclature PCS

Une part importante du questionnement (page 2 du bulletin individuel) permet d'établir des statistiques dans la nomenclature des «Professions et Catégories Socioprofessionnelles» (PCS). Les résultats par profession détaillée sont demandés par de nombreux utilisateurs notamment pour l'analyse de la qualification et des structures professionnelles de la main d'oeuvre. Ils permettent également de produire des statistiques dans les nomenclatures internationales qui nécessitent des tables de passage établies à un niveau fin.

La nomenclature PCS est très détaillée, beaucoup plus que celles d'autres pays et nécessite, outre le libellé, un nombre important de variables annexes (13 au total). Trois de ces variables sont des caractéristiques de l'établissement employeur, la plupart du temps inconnues de l'enquêté. C'est pourquoi on est amené à identifier l'établissement employeur dans le répertoire d'établissements SIRENE sur la base de ses nom et adresse indiqués par le recensé. On peut ainsi accéder aux informations relatives à l'établissement disponibles dans SIRENE, parmi lesquelles l'activité économique, la catégorie juridique et la taille.

Les résultats de la consultation

De nouveaux besoins en données ou des besoins anciens non satisfaits ont été exprimés impliquant pour la plupart l'introduction de nouvelles questions ou des changements de formulation. Un premier examen a permis d'établir la liste des demandes susceptibles d'être prises en compte dans les futurs questionnaires. Divers critères ont influé sur les choix effectués : l'intérêt du thème partagé par de nombreux utilisateurs, la possibilité de définir un questionnement simple et la bonne acceptabilité présumée de ce questionnement par le public, compte tenu, en particulier, de l'expérience acquise lors des recensements antérieurs ou d'enquêtes.

Le tableau ci-après présente les domaines où la consultation a conduit à proposer de nouvelles questions ou formulations de questions.

	FEUILLE DE LOGEMENT	BULLETIN INDIVIDUEL
Nouvelles questions	• **Année d'emménagement** dans le logement • **Surface** du logement • Existence d'un système de **climatisation** • Jouissance d'un **parking** • Jouissance d'une **résidence secondaire** • Equipement en **2 roues** motorisées	• **Année d'arrivée en France** pour les personnes nées à l'étranger • **Niveau de formation** (pour les plus de 14 ans) • **Inscription dans un établissement d'enseignement** pendant l'année scolaire en cours - et **lieu d'études** • **Moyens de transport** utilisés pour le trajet domicile-travail • Occupation d'un **logement occasionnel** (pour repérer les migrants alternants non quotidiens)
Nouvelles formulations impliquant parfois un questionnement supplémentaire	• Année d'achèvement de la construction Prise en compte dans certains cas de l'année de **réhabilitation** au lieu de la date de construction • Statut d'occupation du logement Remplacement de la rubrique «logé gratuitement» par «**autres**» cas pour mieux prendre en compte certaines situations ambiguës • Installations sanitaires Ajout d'un questionnement sur le **nombre de salles d'eau** • Moyen de chauffage Ajout d'une modalité «**tout électrique**» à radiateurs muraux • Equipement en voitures Ajout de la modalité «**2 voitures**»	• **Diplômes** - modification de la grille utilisée au RP90, en particulier ajout de la modalité «**sans diplôme**» - interrogation sur **tous les diplômes** obtenus et non sur le diplôme le plus élevé • **Emploi** - demande de précision sur **le travail à temps partiel** - prise en compte de la situation des «**gérants**» dans la question sur le statut - catégorie professionnelle et fonction principale de l'emploi occupé antérieurement pour les **inactifs** (afin de mieux coder la PCS)

II - Le test de questionnaires

Comme il avait été convenu au début de la consultation, les propositions de nouvelles questions issues de la consultation devaient faire l'objet d'un essai sur le terrain. Il fallait s'assurer que les questions étaient «efficaces» c'est-à-dire que les répondants comprenaient bien le libellé des questions et y répondaient correctement.

C'est ainsi qu'un test de questionnaires s'est déroulé au printemps 1994 dans neuf régions (Aquitaine, Auvergne, Bourgogne, Bretagne, Champagne-Ardenne, Ile-de-France, Languedoc-Roussillon, Limousin, Nord-Pas-de-Calais). Environ 80 communes ont été concernées en totalité ou en partie.

Deux versions de questionnaires testées

Toutes les suggestions ne pouvaient en effet trouver place sur un seul imprimé sans alourdir excessivement le questionnaire et interférer avec la logique des questionnements. De plus, pour certaines questions (nouvelles ou reformulées par rapport au RP90), on souhaitait proposer des formulations alternatives de manière à retenir celles qui s'avéreraient les plus pertinentes.

La répartition des nouvelles questions entre les 2 versions ne s'est pas faite au hasard. Compte tenu du budget alloué au test, l'essai de recensement devait porter sur 30 000 personnes. Il a été décidé de tester sur 15 000 personnes chacune des versions. Cette parité était également respectée au niveau de chaque région. Afin d'optimiser le rendement du test, les 2 versions de questionnaires avaient été mises au point pour être utilisées dans des zones présentant des caractéristiques spécifiques en rapport avec les nouvelles questions à tester. Il fallait en effet que les nouvelles questions soient posées aux personnes les plus susceptibles d'être concernées.

Sans rentrer dans le détail des questions figurant dans l'une ou l'autre version, on peut dire que l'une des versions (A) visait plus particulièrement les banlieues des grandes agglomérations urbaines incluant certains quartiers où la présence d'étrangers était importante. L'année d'entrée en France pour les nés à l'étranger étant une question très délicate, il fallait bien s'assurer qu'elle ait été posée à un nombre suffisant de ressortissants de diverses communautés étrangères. Deux autres questions étaient également posées uniquement dans cette version : la superficie du logement et l'équipement en motos du ménage.

La version B des questionnaires s'adressait à un public plus aisé. Les questions sur la jouissance d'une résidence secondaire et la climatisation y figuraient de même que la question sur l'année de réhabilitation de l'immeuble.

Le déroulement du test

La collecte du test a été réalisée par des enquêteurs ou des agents recenseurs recrutés pour l'occasion. La méthode de collecte utilisée a été celle d'un recensement traditionnel c'est-à-dire le dépôt-retrait. Lors du retrait des questionnaires l'agent recenseur ne devait apporter aucun complément ou modification aux documents qui lui étaient remis. Par contre, il devait à l'aide d'un questionnaire complémentaire interroger le ménage pour obtenir des précisions sur les réponses (ou les non réponses) fournies aux nouvelles questions.

Au total, 14 500 logements ont été recensés dont 12 500 résidences principales totalisant 28 000 habitants.

Les conditions du test différaient de celles d'un recensement réel. L'absence de campagne médiatique autour de l'opération (seules quelques affiches dans les lieux publics et dans les halls d'immeubles), le caractère non obligatoire du test et la non intervention des mairies expliquent les taux de non réponses parfois élevés constatés dans certaines zones.

Par ailleurs, afin de bien cerner les réponses spontanées aux différentes questions, il était demandé aux agents recenseurs, contrairement à ce que l'on préconise dans le cadre d'un recensement, de ne pas intervenir dans le remplissage des questionnaires, même s'ils avaient connaissance de l'information (cas de certaines questions de la feuille de logement). Le questionnaire complémentaire devrait indiquer si l'absence de réponse correspondait à un oubli, un refus ou une mauvaise compréhension de la question.

L'exploitation du test

L'exploitation des résultats du test a reposé sur différents éléments :

- un traitement statistique des documents de collecte (BI, FL) orienté sur l'étude des non réponses,

- un traitement statistique des questionnaires complémentaires pour apprécier la qualité des réponses et tenter de comprendre l'origine des non réponses,

- les bilans qualitatifs des 9 DR ayant participé au test, notamment les rapports des agents recenseurs sur l'accueil et le degré de compréhension des nouvelles questions sur le terrain.

L'effet «test» sur les taux de réponses

L'observation a d'abord porté sur les taux de réponse aux questions qui n'avaient subi aucune modification par rapport au RP90. Dans le bulletin individuel, ces questions avaient trait principalement aux caractéristiques démographiques (sexe, état matrimonial, année et lieu de naissance, lieu de résidence antérieure). Les taux de non réponses enregistrés lors du test se sont révélés voisins de ceux observés en 1990 (entre 1 et 3 % selon les questions) à l'exception de la question sur l'état matrimonial légal où le taux est plus élevé (6 % contre 2 % au RP90).

Par contre, pour les questions figurant sur la feuille du logement (type de logement, nombre de pièces, présence de WC intérieurs, moyen de chauffage, combustible), les taux de non réponses observés lors du test sont nettement supérieurs à ceux du RP90. Alors que ces derniers oscillent entre 1 et 3 % selon les questions, ceux observés au test se situent entre 5 et 9 % et ceci sans tenir compte des ménages qui n'avaient répondu à aucune question de la feuille de logement[1] (environ 5 % des ménages).

Plusieurs éléments peuvent expliquer ces différences dans les taux de non réponses : le caractère non obligatoire du test , la campagne de communication restreinte qui l'a précédé et surtout la non intervention des agents recenseurs dans le remplissage des questionnaires. Ce dernier point est le plus important. Au moment du recensement, beaucoup d'«oublis» des ménages sont aisément réparés par les agents recenseurs à qui on demande de contrôler le bon remplissage des questionnaires. Cela vaut tout particulièrement pour les questions relatives au confort des logements dans les immeubles collectifs où les réponses sont souvent identiques d'un logement à l'autre.

Pour juger du degré d'acceptation des nouvelles questions de la feuille de logement, il a paru nécessaire d'opérer une distinction entre les ménages qui avaient, semble-t-il, répondu assez «sérieusement» au questionnaire et les ménages qui se sont contentés de les «survoler». Deux questions ont alors été privilégiées : le type de logement et le statut d'occupation. Ces questions apparaissent en effet centrales dans l'analyse des conditions de logement et ne sont pas censées poser beaucoup de problèmes de collecte. Les ménages qui ont répondu simultanément à ces deux questions constituaient la première sous-population. Ils étaient en proportion plus nombreux dans la population A (concernée par la version A des questionnaires) que dans la population B (88 % contre 85 %), résultat semble-t-il lié à la structure

---

[1] Certaines des questions de la feuille de logement sont de réponse facile. Les non répondants à l'ensemble de la feuille de logement sont donc des ménages qui n'ont pas fait d'effort particulier pour remplir le questionnaire sans que l'on puisse en tirer un quelconque enseignement sur la difficulté des différentes questions de la feuille de logement.

sociale plus élevée dans l'échantillon B. Bien évidemment, il a été tenu compte, principalement, des taux de réponse de ces ménages, appelés ménages témoins, lorsqu'il s'agissait de juger de l'accueil des nouvelles questions.

Le choix définitif des questions retenues

Au vu des résultats du test, les décisions de maintien ou de suppression des nouvelles questions (ou formulations) se sont imposées facilement dans la majorité des cas. Les questions qui concernaient une population trop restreinte telle que la présence d'un système de climatisation dans le logement, l'équipement en motos ou l'utilisation d'un logement occasionnel ont été écartées (moins de 5 % de personnes concernées parmi les répondants). Il en a été de même lorsque le questionnaire complémentaire avait mis en évidence une mauvaise compréhension du questionnement. C'est ainsi que certains ménages avaient assimilé des travaux d'embellissement du logement (ravalement, peintures...) à une réhabilitation importante dont la date aurait dû se substituer à la date d'achèvement de la construction. Pour la question sur la jouissance d'une résidence secondaire le débat était plus ouvert. Les résultats du test étaient plutôt concluants (4 % de non répondants parmi les ménages témoins et 13 % des répondants concernés) mais les bilans des DR avaient montré que la question avait été jugée parfois «dérangeante». Le Comité Directeur du Recensement a décidé de ne pas retenir la question considérant qu'elle était «à risque» et qu'elle pouvait, par exemple, alimenter le débat lors d'une campagne sur les mal logés.

Une question cependant a fait l'objet d'un «repêchage» en quelque sorte : la superficie du logement. D'une «efficacité» peu probante lors du test (taux de non réponse plutôt élevé : 19 % parmi les ménages témoins - fiabilité incertaine des réponses, d'après les bilans qualitatifs des DR et les sources d'information citées par les répondants lors de l'interview), il semblait, compte tenu de son intérêt, que la question était perfectible si on changeait la formulation. Demander la surface exacte (en m^2) du logement, comme c'était le cas dans le test, provoque parfois quelques réticences par exemple pour des raisons fiscales ou simplement parce que peu de résidents connaissent précisément la superficie de leur logement en particulier en maison individuelle. Il a donc été jugé préférable d'introduire une question sur l'appartenance à une tranche de surface qui évite cet écueil et s'accorde mieux du caractère approximatif des réponses.

Un test de procédures de collecte réalisé à l'automne 1995 a été l'occasion de tester cette nouvelle formulation. Pour contrôler la qualité des réponses, une confrontation de ces dernières avec la surface figurant dans les fichiers de la taxe d'habitation a été effectuée par sondage dans 2 régions (1 700 logements). Pour les deux tiers des résidences principales, il y avait cohérence entre les deux sources d'information. Dans les autres cas, il y avait seulement glissement d'une tranche mais sans biais

systématique. Par contre, le taux de non réponse à la question, quoique plus faible que celui observé avec la première formulation, demeure encore élevé.

Enfin, une demande supplémentaire est parvenue une fois le test des questionnaires effectué. L'Institut Français de l'Environnement (IFEN) appuyé par les ministères de l'Environnement et de l'Agriculture ont demandé d'introduire une question sur l'évacuation et le traitement des eaux usées. Cette question avait déjà été posée de manière simplifiée lors des recensements de 1975 et 1982 mais avait été supprimée en 1990. Compte tenu de l'importance que revêtent aujourd'hui les problèmes d'assainissement et d'épuration des eaux usées, il a été décidé de réintroduire une question sur ce thème dans la feuille de logement.

En définitive, en plus des questions sur la surface et l'assainissement, plusieurs nouvelles questions vont figurer dans les questionnaires du 33ème RP. Ce sera le cas dans la feuille de logement des questions sur l'année d'emménagement, la jouissance d'un parking et le nombre de salles d'eau. Moins de 5 % de non répondants pour chacune de ces 3 questions, bonne qualité des réponses appréciée par le questionnaire complémentaire, bon accueil du public souligné par les bilans qualitatifs.

Dans le bulletin individuel, il s'agit des questions sur l'année d'arrivée en France (pour les personnes nées à l'étranger), le niveau de formation, l'inscription dans un établissement d'enseignement et le lieu d'études, les moyens de transport pour le trajet domicile-travail.

Pour l'année d'arrivée en France, 13 % de personnes nées à l'étranger et non françaises de naissance n'ont pas répondu à la question, mais l'interview complémentaire a montré qu'il s'agissait souvent d'un oubli (11 %) et très rarement d'un refus (2 %). D'autre part, aucune réaction négative n'avait été enregistrée par les agents recenseurs.

Pour les questions sur le niveau de formation et l'inscription dans un établissement d'enseignement, les formulations ainsi que l'emplacement des questions différaient selon les 2 versions de questionnaires. Le test a montré sur ces deux points la meilleure façon de procéder à ce double questionnement. La question sur l'inscription dans un établissement d'enseignement est mieux comprise lorsqu'elle est posée à l'ensemble de la population que lorsqu'elle est formulée de manière à identifier seulement les étudiants. Les réponses aux questions «niveau de formation» et «diplôme» se sont avérées cohérentes. Les enseignements du test ont permis par ailleurs de montrer qu'il était pertinent de supprimer la modalité «aucune scolarité» dans la question sur le niveau de formation et de maintenir la modalité «aucun diplôme» dans la question sur le diplôme, un arbitrage devant être fait entre ces deux modalités, compte tenu des impératifs de place sur le questionnaire. La proportion de gens n'ayant aucun diplôme est en effet plus importante que celle

n'ayant suivi aucune scolarité. De plus, la modalité «aucune scolarité» n'a pas toujours été bien comprise, certaines personnes, d'après les bilans des DR, l'ayant interprétée comme «ne pas avoir fait d'études».

Enfin, la question sur les moyens de transports domicile-travail a été très bien accueillie (moins de 3 % de non répondants) et l'interview complémentaire a permis de détecter quelques erreurs de compréhension auxquelles on a remédié par la suite en modifiant légèrement la formulation.

III - L'incidence de la codification automatique et de la scannérisation éventuelle des questionnaires

Le report de 1997 à 1999 du 33ème RP n'a pas conduit à bouleverser le projet initial. La plupart des décisions déjà prises n'ont pas été remises en cause.

En 1999, les libellés feront l'objet d'une codification automatique. Il est prévu, par ailleurs, de supprimer le poste de préparation des documents avant la saisie où le codage de certains libellés se faisait. Une relecture du bulletin individuel a été opérée de manière à ce que les libellés écrits par les répondants soient compatibles avec les exigences de la codification automatique. En particulier, 2 questions supplémentaires pourraient être posées pour économiser la saisie des libellés des communes-lieu de travail et lieu d'études pour les personnes qui travaillent ou/et sont scolarisées dans leur commune de résidence. En effet, environ 70 % des personnes scolarisées dans un établissement d'enseignement le sont dans leur commune de résidence et la moitié des actifs travaillent dans la commune où ils habitent. Compte tenu des effectifs concernés (15 millions de scolaires ou étudiants, 22 millions d'actifs), l'introduction d'une question permettant de savoir directement, à l'aide d'une case cochée, si la commune lieu d'études ou lieu de travail est la commune de résidence, évite la saisie d'environ 20 millions de libellés de communes, d'où une économie importante. Un test de questionnaires sera effectué pour s'assurer de l'accueil de ces nouvelles formulations.

Le recueil et la saisie des informations pourraient aussi être sensiblement modifiés par l'introduction d'une méthode de scannérisation. Une étude de faisabilité de ce type de méthode pour le recensement a été confiée à un intervenant extérieur. Des matériels et logiciels sont, en effet, actuellement opérationnels pour acquérir les images des questionnaires et reconnaître la totalité des cases cochées et la quasi-totalité des chiffres précasés. Quant aux libellés manuscrits, ils sont présentés sur l'écran d'un poste de reprise et saisis manuellement mais sans avoir recours au document papier. Il n'y a plus qu'une seule manipulation de documents, alors que la saisie de l'échantillon au ¼, défini a posteriori, étant disjointe de celle traitée

exhaustivement, nécessitait une seconde manipulation[1] . Cette nouvelle technologie a été expérimentée avec succès pour la saisie des formulaires du dernier recensement israélien (1995).

Les conclusions de l'étude ont été rendues en septembre. Il s'avère que dans le cadre du recensement français, cette méthode de scannérisation est réalisable techniquement et financièrement intéressante. Si cette méthode était retenue, des modifications dans la présentation des documents seraient alors nécessaires (fond blanc obligatoire pour les imprimés, doublement de la surface des cases à cocher, inscription des chiffres dans les cases). Il conviendrait de s'assurer lors d'un prochain test de questionnaires qu'elles n'ont pas d'effets négatifs sur les réponses des recensés.

---

[1] Au RP90, les documents ont fait l'objet d'une double manipulation :
1 - Envoi des documents (600 tonnes environ) des DR aux lieux de saisie puis retour en DR
2 - Après tirage de l'échantillon au 1/4, reprise de la totalité des dcouments en DR pour une saisie-codification «en ligne» des logements sondés.
La scannérisation évite cette seconde manipulation.

ANNEXES

RECOMMANDATIONS DE L'ONU

Caractéristiques fondamentales	Caractéristiques supplémentaires
Caractéristiques géographiques des personnes	
1. Lieu de résidence habituelle	1. Lieu de présence au moment du recensement 2. Résidence dans une exploitation agricole ou non
Caractéristiques dérivées a) Population totale b) Localité 2. Lieu de résidence à une date	Caractéristiques dérivées a) Zones urbaines et rurales 3. Durée de la résidence de référence antérieure 4. Lieu de résidence antérieure 5. Année (ou période) d'immigration dans le pays
Caractéristiques démographiques des personnes	
3. Sexe 4. Age 5. Situation matrimoniale (légale)	6. Situation matrimoniale (de fait) (donnée primaire ou dérivée)
6. Pays de naissance et/ou de citoyenneté (nationalité juridique)	7. Lieu de naissance (population née dans le pays) 8. Groupe national et/ou ethnique 9. Langue 10. Religion 11. Nombre total d'enfants nés vivants i) des femmes mariées et ii) si possible, de toutes les femmes 12. Date du i) premier mariage et ii) du mariage actuel de la femme
Caractéristiques économiques des personnes	
8. Type d'activité (effective ou habituelle)	13. Type d'activité (habituel ou effectif) 14. Temps de travail 15. Durée du chômage 16. Principal moyen d'existence 17. Revenu 18. Liens de dépendance (réels ou présumés)
8. Profession 9. Branche d'activité économique 10. Situation d'activité économique (employeur, salarié, etc.)	19. Profession secondaire 20. Secteur d'emploi 21. Nombre de personnes employées par l'employeur

Caractéristiques fondamentales	Caractéristiques supplémentaires
Caractéristiques dérivées	
c) Groupe socio-économique 11. Lieu de travail	22. Lieu de l'école, de l'université, etc. 23. Trajet jusqu'au lieu de travail
Caractéristiques d'instruction de la personne	
12. Niveau d'instruction	24. Diplômes obtenus 25. Fréquentation scolaire 26. Aptitude à lire et à écrire
Caractéristiques du ménage et de la famille de la personne	
13. Lien avec la personne de référence du ménage privé	27. Type du ménage institutionnel de l'établissement collectif dans dans lequel la personne vit 28. La personne est-elle pensionnaire d'un ménage institutionnel ou d'un établissement collectif ?
Caractéristiques dérivées	Caractéristiques dérivées
d) Position dans le ménage e) Position dans la famille	b) Position dans la famille élargie
Caractéristiques du noyau familial	
Caractéristiques dérivées	Caractéristiques dérivées
f) Type de noyau familial g) Taille du noyau familial h) Nombre d'enfants au-dessous d'un âge déterminé i) Nombre de membres actifs	c) Type de famille élargie d) Groupes d'âge déterminés des enfants e) Nombre de membres dont le principal moyen d'existence est une activité économique f) Nombre de membres qui sont des personnes à charge
Caractéristiques des ménages privés	
Caractéristiques dérivées	Caractéristiques dérivées
j) Type de ménage privé k) Taille du ménage privé l) Nombre de membres actifs	g) Composition des ménages privés par génération h) Nombre de membres dont le principal moyen d'existence est une activité économique i) Nombre de membres qui sont des personnes à charge

Caractéristiques fondamentales	Caractéristiques supplémentaires
Caractéristiques dérivées	
m) Nombre d'enfants au-dessous d'un âge déterminé	
n) Nombre de membres à l'âge de la retraite	
14. Modalités de jouissance du logement	29. Ménages vivant seuls dans un logement ou partageant un logement
	30. Loyer
	31. Biens de consommation durables appartenant au ménage
	32. Nombre de voitures automobiles par ménage
Caractéristiques des unités d'habitation et autres locaux d'habitation	
15. Régime de propriété	
16. Emplacement du local d'habitation	
17. Type de local d'habitation	
18. Régime d'occupation	33. Type de non-occupation
19. Occupation par un ou plusieurs ménages	
20. Nombre d'occupants	
21. Nombre de pièces	34. Surface utile et/ou habitable
22. Cuisine	35. Installations pour la préparation des aliments
23. Système d'adduction d'eau	36. Eau chaude
24. Lieux d'aisances	37. Type de système d'évacuation des eaux usées
25. Salles d'eau	
26. Type de chauffage	38. Principale source d'énergie pour le chauffage
	39. Isolation
	40. Electricité
	41. Gaz sur réseau de distribution
	42. Téléphone
	43. Emplacement du logement dans le bâtiment
Caractéristiques des bâtiments comportant des logements	
27. Type de bâtiment	44. Nombre d'étages
	45. Nombre de logements dans le bâtiment
	46. Le bâtiment contenant le logement est-il ou non un bâtiment agricole ?
	47. Ascenseur
28. Epoque de construction	48. Matériaux de construction de parties déterminées du bâtiment

FEUILLE DE LOGEMENT

IMPRIMÉ NUMÉRO **1**

À remplir pour tout logement d'habitation occupé ou non.

Cadre à remplir par l'agent recenseur :

- Numéro du district de recensement
- Rang de l'immeuble dans le district
- Rang du logement dans l'immeuble

Cachet de la mairie

RECENSEMENT DE LA POPULATION

1990

RÉPUBLIQUE FRANÇAISE

Remplissez les quatre pages très lisiblement.

(1) **NOM ET PRÉNOM DE L'OCCUPANT :** _____

(2) **ADRESSE DU LOGEMENT :** N° _____ Rue (ou lieudit) : _____

_____

Commune : _____ Département : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

(3) **LOCALISATION DU LOGEMENT DANS L'IMMEUBLE** (SI CE DERNIER COMPREND DEUX LOGEMENTS OU PLUS)

a. Escalier : _____ *b.* Étage : _____

c. Situation sur le palier : _____

_____ Si la porte d'entrée du logement a un numéro, indiquez ce numéro : _____

(4) **CATÉGORIE DE LOGEMENT**

1 ☐ **Résidence principale** (logement ou pièce indépendante où le ménage demeure la plus grande partie de l'année). Une chambre louée par un étudiant sera sa résidence principale. ⇨ Remplir un bulletin individuel pour chaque personne inscrite en liste 1, page 2.

2 ☐ **Logement** (ou pièce indépendante) **utilisé occasionnellement.**

3 ☐ **Résidence secondaire** ou logement loué (ou à louer) pour les loisirs ou les vacances. ⇨ Pas de bulletin individuel sauf si des personnes habitent le logement au moment du recensement. Dans ce cas, établir les bulletins individuels correspondants, en prenant soin de remplir le cadre «PERSONNES DE PASSAGE». Ne pas remplir les pages 2 et 3 de la feuille de logement.

4 ☐ **Logement vacant** (sans occupant, disponible ou non pour la vente ou la location). ⇨ Pas de bulletin individuel. Ne pas remplir les pages 2 et 3 de la feuille de logement.

5 ☐ **Caravane, habitation mobile.** ⇨ Remplir un bulletin individuel pour chaque personne inscrite en liste 1. Ne pas remplir la page 4 ; à classer à part.

RÉCAPITULATION GÉNÉRALE

À remplir par l'agent recenseur	*À remplir par la mairie*			
(5) Nombre d'imprimés n° 2 recueillis ***(liste 1 seulement)***	(6) Nombre d'imprimés n° 2 supprimés *(faisant double emploi avec des imprimés n° 5 réintégrés)*	(7) Nombre d'imprimés n° 2 ajoutés *(bulletins reçus pour des personnes en déplacement et ne faisant pas double emploi)*	(8) Nombre d'imprimés n° 5 réintégrés	(9) Population après réintégrations (9) = (5) − (6) + (7) + (8)

Visa n° 90X001 Ec. du ministre d'État, ministre de l'Économie, des Finances et du Budget. Enquête statistique obligatoire (loi n° 51-711 du 7 juin 1951 modifiée). Les questionnaires, collectés par l'intermédiaire des mairies, sont exclusivement destinés à l'INSEE. Il est interdit de prendre copie du présent questionnaire sous peine des sanctions prévues à l'article 44 de la loi n° 78-17 du 6 janvier 1978. Les données anonymes résultant de l'exploitation du recensement peuvent être cédées par l'INSEE à d'autres organismes (voir notice d'information).

PAGE 1

CARACTÉRISTIQUES GÉNÉRALES DE L'IMMEUBLE ET DU LOGEMENT

(1) TYPE DE LOGEMENT

- Logement-foyer pour personne(s) âgée(s) ..... ☐ 1
- Ferme ..... ☐ 2
- Chambre d'hôtel ..... ☐ 3
- Construction provisoire, habitation de fortune ..... ☐ 4
- Pièce indépendante (ayant sa propre entrée) ..... ☐ 5
- Maison individuelle ..... ☐ 6
- Logement dans un immeuble collectif ..... ☐ 7
- Logement dans un immeuble à usage autre que d'habitation *(usine, atelier, immeuble de bureaux, magasin, école, collège, hôpital, mairie, gare, bureau de poste, stade, etc.)* ..... ☐ 8

(2) ANNÉE D'ACHÈVEMENT DE LA CONSTRUCTION DE LA MAISON OU DE L'IMMEUBLE

- Avant 1915 ..... ☐ 1
- de 1915 à 1948 ..... ☐ 2
- de 1949 à 1967 ..... ☐ 3
- de 1968 à 1974 ..... ☐ 4
- de 1975 à 1981 ..... ☐ 5
- 1982 ou après ..... ☐ 6
- Immeuble en cours de construction partiellement habité ..... ☐ 7

Si les différentes parties ne sont pas de la même époque, indiquez l'année d'achèvement de la partie habitée, ou de la partie habitée la plus importante.

(3) NOMBRE DE PIÈCES D'HABITATION ⇨ ☐

- Comptez les **pièces telles que** chambres à coucher, salle à manger, salle de séjour, quelle que soit leur surface.
- Ne comptez la cuisine que si sa surface est supérieure à 12 m².
- Ne comptez pas les **pièces telles que** couloir, salle de bains, W.-C., buanderie, etc.
- Ne comptez pas les pièces à usage exclusivement professionnel *(exemples : cabinet de médecin, atelier d'artisan, etc.)*.

(4) ÊTES-VOUS ?

- Propriétaire du logement *(y compris les différentes formes d'accession à la propriété)* ..... ☐ 1
- Locataire du logement par exemple par des parents, des amis ou toute autre personne *(y compris le cas des personnes qui occupent gratuitement qu'elles ont vendu ou légué ou dont elles ont la jouissance par héritage)* ..... ☐ 2
- Locataire ou sous-locataire d'un logement loué meublé ..... ☐ 3
- Locataire ou sous-locataire d'un logement loué meublé ou d'une chambre d'hôtel ..... ☐ 4

(5) LE LOGEMENT APPARTIENT-IL À UN ORGANISME HLM ? OUI ☐ 1 NON ☐ 2

(6) INSTALLATIONS SANITAIRES

Avez-vous une baignoire ou une douche ?

- Baignoire ..... ☐ 1
- Douche seulement ..... ☐ 2
- Ni baignoire ni douche ..... ☐ 3

(7) W.-C.

Sont-ils situés à l'intérieur du logement ? OUI ☐ 1 NON ☐ 2

(8) MOYEN DE CHAUFFAGE DU LOGEMENT

- **Chauffage central collectif** *(commun à la totalité ou à la plupart des logements de l'immeuble, y compris le chauffage urbain)* ..... ☐ 1
- **Chauffage central individuel** avec une chaudière propre au logement *(y compris pompe à chaleur et chauffage électrique à radiateurs muraux)* ..... ☐ 2
- **Autres moyens de chauffage** *(poêle, cheminée, cuisinière, radiateurs mobiles, appareils à accumulation, etc.)* ..... ☐ 3

(9) COMBUSTIBLE PRINCIPAL UTILISÉ POUR LE CHAUFFAGE (quel que soit le moyen de chauffage du logement)

Cochez une seule case.

- Chauffage urbain ..... ☐ 1
- Gaz de ville ou de réseau ..... ☐ 2
- Fioul (mazout) ..... ☐ 3
- Électricité ..... ☐ 4
- Gaz en bouteilles ou citerne ..... ☐ 5
- Charbon ..... ☐ 6
- Bois ..... ☐ 7

(10) NOMBRE DE VOITURES DONT DISPOSENT LES HABITANTS DU LOGEMENT

- Aucune ..... ☐ 0
- 1 ..... ☐ 1
- 2 ou plus ..... ☐ 2

(11) CE LOGEMENT EST-IL LE SIÈGE D'UNE EXPLOITATION AGRICOLE ? OUI ☐ 1 NON ☐ 2

Si oui :

(a) Superficie agricole utilisée :

_____ **hectares,** _____ **ares**

Ne tenez pas compte des bois, étangs, terrains à bâtir, parcs et jardins d'agrément, landes et friches improductives, bâtiments et cours.

(b) Orientation des productions agricoles

Cochez une seule case.

- Exploitation avec une production principale
  - Polyculture (cultures de terres labourables) ..... ☐ 1
  - Maraîchage ou horticulture ..... ☐ 2
  - Vigne ou arbres fruitiers ..... ☐ 3
  - Élevage d'herbivores (bovins, ovins...) ..... ☐ 4
  - Élevage de granivores (porcs, volailles...) ..... ☐ 5
- Autres cas
  - Polyculture - élevage ..... ☐ 6
  - Élevage d'herbivores et de granivores ..... ☐ 7
  - Autres ..... ☐ 8

IMPRIMERIE NATIONALE 9 917 201 057

BULLETIN INDIVIDUEL

RECENSEMENT DE LA POPULATION

1990

RÉPUBLIQUE FRANÇAISE

IMPRIMÉ NUMÉRO **2**

Cadre à remplir par l'INSEE : ☐1 ☐2 ☐3

Cachet de la mairie

(1) NOM, Prénoms
Écrivez le nom en capitales
(exemple : ALLARD, épouse MAURIN, Françoise).

(2) ADRESSE

(3) SEXE
Masculin ☐ 1
Féminin ☐ 2

(4) ÉTAT MATRIMONIAL LÉGAL
Une personne vivant en union libre cochera la case correspondant à sa situation juridique présente : si, par exemple, elle est célibataire, elle cochera la 1re case.
Une personne en instance de divorce, ou séparée de son conjoint légitime, cochera la 2e case.
Célibataire ☐ 1
Marié(e) ou remarié(e) ☐ 2
Veuf(ve) ☐ 3
Divorcé(e) ☐ 4

(5) DATE ET LIEU DE NAISSANCE
Né(e) le : _____
(jour, mois, année)
à (commune) : _____
Département : _____
(pays pour l'étranger, territoire pour les TOM)

(6) NATIONALITÉ
- Français de naissance (y compris par réintégration) ☐ 1
- Devenu français par naturalisation, mariage, déclaration ou à votre majorité ☐ 2
  Indiquez votre nationalité antérieure : _____
- Étranger ☐ 3
  Indiquez votre nationalité : _____

(7) OÙ HABITIEZ-VOUS LE 1er JANVIER 1982 ?
(pour toute personne née avant le 1er janvier 1982)
Si, le 1er janvier 1982, vous étiez militaire ou élève interne, indiquez l'adresse de votre résidence personnelle à cette date et non celle de l'établissement (caserne, internat).
- Dans le même logement que maintenant ☐ 1
- Dans un autre logement de la même commune ☐ 2
  (ou du même arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une autre commune ☐ 3
  (ou un autre arrondissement s'il s'agit de Paris, Lyon, Marseille)

  Indiquez cette autre commune :
  Commune : _____
  (pour Paris, Lyon, Marseille, précisez l'arrondissement)
  Département : _____
  (pays pour l'étranger, territoire pour les TOM)

LES QUESTIONS 8 À 21 NE S'ADRESSENT QU'AUX PERSONNES DE 14 ANS OU PLUS

(8) INDIQUEZ VOTRE DIPLÔME LE PLUS ÉLEVÉ
- Certificat d'études primaires ☐ 1
- BEPC, brevet élémentaire, brevet des collèges ☐ 2
- CAP ☐ 3
- BEP ☐ 4
- Baccalauréat, brevet professionnel ou de technicien, autre brevet (BEA, BEC, BEI, etc.) ☐ 5
- Diplôme universitaire de 1er cycle, BTS, DEST, DUT, diplôme des professions sociales ou de la santé ☐ 6
- Diplôme universitaire de 2e ou 3e cycle, diplôme d'ingénieur, d'une grande école, etc. ☐ 7

Si vous travaillez, *passez au verso (questions 12 à 21)*
y compris :
- *Si vous êtes en congé de maladie ou de maternité ;*
- *Si vous aidez un membre de votre famille dans son travail même sans être rémunéré ;*
- *Si vous êtes apprenti sous contrat, stagiaire rémunéré (TUC, SIVP...), etc.*

Si vous ne travaillez pas (ou plus),
répondez aux questions 9 à 11

(9) ÊTES-VOUS ?
- Élève, étudiant, stagiaire non rémunéré ☐ 1
- Chômeur *(inscrit ou non à l'ANPE)* ☐ 2
- Retraité *(ancien salarié)* ou préretraité ☐ 3
- Retiré des affaires *(ancien agriculteur, ancien artisan, ancien commerçant, etc.)* ☐ 4
- Femme au foyer ☐ 5
- Autre inactif *(y compris les personnes ne percevant qu'une pension de réversion)* ☐ 6

(10) AVEZ-VOUS DÉJÀ TRAVAILLÉ ?
OUI ☐ 1 → Quelle était votre profession principale ?
NON ☐ 2

(11) CHERCHEZ-VOUS UN EMPLOI ?
- Vous ne cherchez pas d'emploi ☐ 1
- Vous cherchez un emploi depuis :
  - moins de 3 mois ☐ 2
  - 3 mois à moins de 1 an ☐ 3
  - 1 an à moins de 2 ans ☐ 4
  - 2 ans ou plus ☐ 5

Signez le bulletin au bas de la page 2

POUR LES PERSONNES DE PASSAGE (voir page 3 de l'imprimé n° 1), adresse de la résidence habituelle :
N° _____ Rue (ou lieudit) : _____ Commune : _____ Département : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

IMPRIMERIE NATIONALE N° 172 204 064

VOUS TRAVAILLEZ

(12) PROFESSION EXERCÉE ACTUELLEMENT
Soyez très précis. Exemples : ouvrier électricien d'entretien, chauffeur de poids lourds, vendeur en électroménager, ingénieur chimiste, caissière de libre-service, etc.

(13) AIDEZ-VOUS UN MEMBRE DE VOTRE FAMILLE DANS SON TRAVAIL ?
(Exploitation agricole ou artisanale, commerce, profession libérale, etc.)
OUI ☐ 1
NON ☐ 2

(14) Si vous êtes agent de l'État, d'une collectivité locale, d'un hôpital public, d'un service public (EDF, SNCF, etc.) **ou militaire de carrière, PRÉCISEZ VOTRE CLASSIFICATION** *(corps, grade, etc.)*

(15) OÙ TRAVAILLEZ-VOUS ?

(a) Adresse de votre lieu de travail :

N° _____ Rue ou lieudit : _____

Commune : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

Département : ☐☐ _____
(ou pays)

(b) Nom (ou raison sociale) de l'établissement qui vous emploie ou que vous dirigez :

(c) Activité de cet établissement :
Soyez très précis. *Exemples : commerce de vin* ***en gros,*** *fabrication de charpentes* ***métalliques,*** *transport routier* ***de voyageurs,*** *etc.*

(d) Adresse de cet établissement, si elle est différente de celle déclarée à la question 15 a.

(16) ÊTES-VOUS ?
- Salarié ☐ 1
- À votre compte ☐ 2
(exploitant agricole, artisan, commerçant, industriel, profession libérale, aide familial non salarié, etc.)

(17) TRAVAILLEZ-VOUS ?
- À temps partiel ☐ 1
- À temps complet ☐ 2

(18) SI VOUS ÊTES À VOTRE COMPTE
Combien de salariés employez-vous ?
Ne comptez ni les apprentis ni les gens de maison. Dans l'agriculture, comptez seulement les salariés permanents.
- Aucun ☐ 1
- 1 ou 2 ☐ 2
- 3 à 9 ☐ 3
- 10 ou plus ☐ 4

LES QUESTIONS 19 À 21 NE S'ADRESSENT QU'AUX SALARIÉS ET AUX STAGIAIRES RÉMUNÉRÉS.

(19) Si vous êtes dans l'une des situations suivantes, cochez la case correspondant à votre cas :
- Apprenti sous contrat ☐ 1
- Exerçant un travail d'utilité collective (TUC, etc.) ☐ 2
- Sous contrat d'adaptation ou de qualification ☐ 3
- Stagiaire (principalement en entreprise : SIVP, etc.) ☐ 4
- Stagiaire (principalement dans un centre de formation : FPA, etc.) ☐ 5
- Placé par une agence d'intérim ☐ 6
- Sous contrat de travail à durée déterminée ☐ 7

(20) INDIQUEZ LA POSITION PROFESSIONNELLE DE VOTRE EMPLOI ACTUEL :
- Manœuvre ou ouvrier spécialisé (OS1, OS2, OS3, etc.) ☐ 1
- Ouvrier qualifié ou hautement qualifié (P1, P2, P3, TA, OQ, etc.) ☐ 2
- Agent de maîtrise dirigeant des ouvriers, maîtrise administrative ou commerciale ☐ 3
- Agent de maîtrise dirigeant des techniciens ou d'autres agents de maîtrise ☐ 4
- Technicien, dessinateur, VRP (non cadre) ☐ 5
- Instituteur, assistant(e) social(e), infirmier(e) et personnel de catégorie B de la fonction publique ☐ 6
- Ingénieur ou cadre *(les employés, techniciens, agents de maîtrise n'ayant pas la qualité de cadre ne devront pas se classer ici, même s'ils cotisent à une caisse de retraite des cadres)* ☐ 7
- Professeur et personnel de catégorie A de la fonction publique ☐ 8
- Employé de bureau, employé de commerce, agent de service, aide soignant(e), gardienne d'enfants, personnel de catégorie C ou D de la fonction publique ☐ 9
- Autre cas. *Précisez :* _____ ☐ 0

(21) QUELLE EST VOTRE FONCTION PRINCIPALE ?
- Production, fabrication, chantiers ☐ 1
- Installation, entretien, réglage, réparation ☐ 2
- Nettoyage, gardiennage, travail ménager ☐ 3
- Manutention, magasinage, transports ☐ 4
- Secrétariat, saisie, guichet, standard ☐ 5
- Gestion, comptabilité, fonctions administratives ☐ 6
- Commerce, vente, technico-commercial ☐ 7
- Recherche, études, méthodes, informatique ☐ 8
- Directeur général ou un de ses adjoints directs ☐ 9
- Autre cas ☐ 0
Précisez (enseignement, santé, information, etc.) :

Visa n° 90X001 Ec, du ministre d'État, ministre de l'Économie, des Finances et du Budget. Enquête statistique obligatoire (loi n° 51-711 du 7 juin 1951 modifiée). Les questionnaires, collectés par l'intermédiaire des mairies, sont exclusivement destinés à l'INSEE. Il est interdit de prendre copie du présent questionnaire sous peine des sanctions prévues à l'article 44 de la loi n° 78-17 du [illegible] vier 1978. Les données anonymes résultant de l'exploitation du recensement peuvent être c par l'INSEE à d'autres organismes (voir notice d'information).

À _____, le _____ 1990
Signature du déclarant :

PAGE 3

FEUILLE DE LOGEMENT

RECENSEMENT DE LA POPULATION

1994

IMPRIMÉ NUMÉRO **1A**

À remplir pour tout logement d'habitation occupé ou non

Cadre à remplir par l'agent recenseur

Numéro du district de recensement

Rang de l'immeuble dans le district

Rang du logement dans l'immeuble

APPLICATION PILOTE A

Remplissez les quatre pages très lisiblement.

(1) **NOM ET PRÉNOM DE L'OCCUPANT :** _____

(2) **ADRESSE DU LOGEMENT :** N° _____ Rue (ou lieudit) : _____

_____

Commune : _____ Département : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

(3) **LOCALISATION DU LOGEMENT DANS L'IMMEUBLE** (SI CE DERNIER COMPREND DEUX LOGEMENTS OU PLUS)

a. Escalier : _____ b. Étage : _____

c. Situation sur le palier : _____

_____ Si la porte d'entrée du logement a un numéro, indiquez ce numéro : _____

(4) **CATÉGORIE DE LOGEMENT**

1 ☐ **Résidence principale** (logement ou pièce indépendante où le ménage demeure la plus grande partie de l'année). Une chambre louée par un étudiant sera sa résidence principale. → Remplir un bulletin individuel pour chaque personne inscrite en liste 1, page 2.

2 ☐ **Logement** (ou pièce indépendante) **utilisé occasionnellement.**

3 ☐ **Résidence secondaire** ou logement loué (ou à louer) pour les loisirs ou les vacances. → Pas de bulletin individuel. Ne pas remplir les pages 2 et 3 de la feuille de logement.

4 ☐ **Logement vacant** (sans occupant, disponible ou non pour la vente ou la location). → Pas de bulletin individuel. Ne pas remplir les pages 2 et 3 de la feuille de logement.

RÉCAPITULATION GÉNÉRALE

à remplir par l'agent recenseur

(5)

Nombre d'imprimés n° 2 recueillis *(liste 1 seulement)* ☐

Enquête statistique non obligatoire.
Questionnaire confidentiel destiné à l'INSEE.
La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés garantit aux individus un droit d'accès et de rectification pour les informations les concernant. Ce droit d'accès peut être exercé pendant le délai où ces informations sont gardées sous forme nominative auprès des directions régionales de l'INSEE.

CARACTÉRISTIQUES GÉNÉRALES DE L'IMMEUBLE ET DU LOGEMENT

(1) TYPE DE LOGEMENT

- Logement-foyer pour personne(s) âgée(s) ☐ 1
- Chambre d'hôtel ☐ 2
- Construction provisoire, habitation de fortune ☐ 3
- Pièce indépendante (ayant sa propre entrée) ☐ 4
- Maison individuelle (y compris avec local professionnel), ferme ☐ 5
- Logement dans un immeuble collectif ☐ 6
- Logement dans un immeuble à usage autre que d'habitation *(usine, atelier, immeuble de bureaux, magasin, école, collège, hôpital, mairie, gare, bureau de poste, stade, etc.)* ☐ 7

(2) ANNÉE D'ACHÈVEMENT DE LA CONSTRUCTION DE LA MAISON OU DE L'IMMEUBLE

- Avant 1915 ☐ 1
- de 1915 à 1948 ☐ 2
- de 1949 à 1967 ☐ 3
- de 1968 à 1974 ☐ 4
- de 1975 à 1981 ☐ 5
- de 1982 à 1989 ☐ 6
- 1990 ou après ☐ 7
- Immeuble en cours de construction partiellement habité ☐ 8

Si les différentes parties ne sont pas de la même époque, indiquez l'année d'achèvement de la partie habitée ou de la partie habitée la plus importante.

(3) NOMBRE DE PIÈCES D'HABITATION → ☐☐

- Comptez les pièces telles que chambres à coucher, salle à manger, salle de séjour, quelle que soit leur surface.
- Ne comptez la cuisine que si sa surface est supérieure à 12 m².
- Ne comptez pas les pièces telles que couloir, salle de bains, W.-C., buanderie, etc.
- Ne comptez pas les pièces à usage exclusivement professionnel *(exemples : cabinet de médecin, atelier d'artisan, etc.)*.

(4) SURFACE DU LOGEMENT → ☐☐☐ mètres carrés

- Tenez compte de toutes les pièces, y compris couloir, cuisine, salle de bains, W.-C., etc.
- Ne tenez pas compte des pièces à usage exclusivement professionnel.
- Ne tenez pas compte des balcons, terrasses, vérandas, caves, parkings, greniers.

(5) ÊTES-VOUS ?

- Propriétaire du logement *(y compris les différentes formes d'accession à la propriété)* ☐ 1
- Locataire ou sous-locataire d'un logement loué vide. ☐ 2
- Locataire ou sous-locataire d'un logement loué meublé ou d'une chambre d'hôtel ☐ 3
- Logé gratuitement, par exemple par des parents, des amis, ou votre employeur *(y compris le cas des personnes occupant un logement qu'elles ont vendu en viager ou dont elles ont la jouissance par usufruit)* ☐ 4

(6) EN QUELLE ANNÉE AVEZ-VOUS EMMÉNAGÉ DANS CE LOGEMENT ? → ☐☐☐☐

(Si tous les occupants actuels du logement ne sont pas arrivés en même temps, indiquez la date d'emménagement du premier arrivé. Si cette personne a toujours vécu dans le logement, reportez sa date de naissance.)

(7) LE LOGEMENT APPARTIENT-IL À UN ORGANISME HLM ? (office, société, ou OPAC) OUI ☐ 1 NON ☐ 2

(8) AVEZ-VOUS UNE BAIGNOIRE OU UNE DOUCHE ? OUI ☐ 1 NON ☐ 2

(9) NOMBRE DE SALLES D'EAU, *pièces réservées à la toilette contenant au moins un lavabo et une baignoire (ou une douche)*

- Aucune ☐ 0
- 1 ☐ 1
- 2 ou plus ☐ 2

(10) W.-C.

Sont-ils situés à l'intérieur du logement ? OUI ☐ 1 NON ☐ 2

(11) MOYEN DE CHAUFFAGE DU LOGEMENT

- **Chauffage central collectif** *(commun à la totalité ou à la plupart des logements de l'immeuble, y compris le chauffage urbain)* ☐ 1
- **Chauffage central individuel** avec une chaudière propre au logement ☐ 2
- **Chauffage « tout électrique »** à radiateurs muraux ☐ 3
- **Autres moyens de chauffage** *(poêle, cheminée, cuisinière, radiateurs mobiles, appareils à accumulation, etc.)* ☐ 4

(12) COMBUSTIBLE PRINCIPAL UTILISÉ POUR LE CHAUFFAGE (quel que soit le moyen de chauffage du logement)

Cochez une seule case.

- Chauffage urbain ☐ 1
- Gaz de ville ou de réseau ☐ 2
- Fioul (mazout) ☐ 3
- Électricité ☐ 4
- Gaz en bouteilles ou citerne ☐ 5
- Charbon ☐ 6
- Bois ☐ 7

(13) GARAGE - BOX - PARKING

Disposez-vous, pour votre usage personnel, d'un emplacement réservé de stationnement situé dans l'immeuble ou la propriété ? OUI ☐ 1 NON ☐ 2

(14) NOMBRE DE VOITURES DONT DISPOSENT LES HABITANTS DU LOGEMENT

- Aucune ☐ 0
- 1 ☐ 1
- 2 ☐ 2
- 3 ou plus ☐ 3

(15) AU MOINS UN DES HABITANTS DU LOGEMENT DISPOSE-T-IL D'UNE MOTO ? (125 cm³ ou plus) OUI ☐ 1 NON ☐ 2

(16) CE LOGEMENT EST-IL LE SIÈGE D'UNE EXPLOITATION AGRICOLE ? OUI ☐ 1 NON ☐ 2

Si oui :

Superficie agricole utilisée :

_____ **hectares,** _____ **ares**

Ne tenez pas compte des bois, étangs, terrains à bâtir, parcs et jardins d'agrément, landes et friches improductives, bâtiments et cours.

PAGE 4

BULLETIN INDIVIDUEL

RECENSEMENT DE LA POPULATION

1994

IMPRIMÉ NUMÉRO **2A**

Remplissez lisiblement. Signez au bas de la page 2.

Cadre à remplir par l'INSEE

APPLICATION PILOTE A

LES QUESTIONS 9 À 26 NE S'ADRESSENT QU'AUX PERSONNES DE 14 ANS OU PLUS

(1) NOM, Prénoms

Écrivez le nom en capitales
(exemple : ALLARD, épouse MAURIN, Françoise).

(2) ADRESSE

(3) SEXE

Masculin ☐ 1
Féminin ☐ 2

(4) ÉTAT MATRIMONIAL LÉGAL

Une personne vivant en union libre cochera la case correspondant à sa situation juridique présente : si, par exemple, elle est célibataire, elle cochera la 1re case.

Une personne en instance de divorce, ou séparée de son conjoint légitime, cochera la 2e case.

Célibataire ☐ 1
Marié(e) ou remarié(e) ☐ 2
Veuf(ve) ☐ 3
Divorcé(e) ☐ 4

(5) DATE ET LIEU DE NAISSANCE

Né(e) le : _____
(jour, mois, année)

à (commune) : _____

Département : _____
(pays pour l'étranger, territoire pour les TOM)

Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ? _____

(6) NATIONALITÉ

- **Français de naissance** (y compris par réintégration) ☐ 1
- **Devenu français** ☐ 2
  Indiquez votre **nationalité antérieure :**
  _____
- **Étranger** ☐ 3
  Indiquez votre **nationalité :**
  _____

(7) ÊTES-VOUS INSCRIT pour l'année scolaire 1993-1994 dans un établissement d'enseignement ?

OUI ☐ 1
NON ☐ 2

Si oui, précisez le lieu d'études :

Commune : _____ *Département :* ☐☐

(8) OÙ HABITIEZ-VOUS LE 1er JANVIER 1987 ?

(pour toute personne née avant le 1er janvier 1987)

Si, le 1er janvier 1987, vous étiez militaire ou élève interne, indiquez l'adresse de votre résidence personnelle à cette date et non celle de l'établissement (caserne, internat).

- Dans le **même logement** que maintenant ☐ 1
- Dans un **autre logement** de la **même commune** ☐ 2
  (du même arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une **autre commune** ☐ 3
  (ou un autre arrondissement s'il s'agit de Paris, Lyon, Marseille)

Indiquez cette **autre commune :**

Commune : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

Département : _____
(pays pour l'étranger, territoire pour les TOM)

(9) QUEL NIVEAU D'ÉTUDES AVEZ-VOUS ATTEINT ?

- Enseignement primaire ☐ 1
- Collège, classes de 6e à 3e, préparation CAP, BEP ☐ 2
- Classe de seconde, première ou terminale ☐ 3
- Enseignement supérieur ☐ 4

(10) INDIQUEZ VOTRE DIPLÔME LE PLUS ÉLEVÉ

- Aucun diplôme ☐ 0
- Certificat d'études primaires ☐ 1
- BEPC, brevet élémentaire, brevet des collèges ☐ 2
- CAP ☐ 3
- BEP ☐ 4
- Baccalauréat général (Philo., Sciences Ex., Math. élém., A, B, C, D, E, L, S, ES...) ☐ 5
- Baccalauréat technologique (F, G, H, STI...) ou professionnel, brevet professionnel ou de technicien, autre brevet (BEA, BEC, BEI, etc.), capacité en droit ☐ 6
- Diplôme universitaire de 1er cycle, BTS, DEST, DUT, diplôme des professions sociales ou de la santé ☐ 7
- Diplôme universitaire de 2e ou 3e cycle, diplôme d'ingénieur, d'une grande école, etc... ☐ 8

(11) QUELLE EST VOTRE SITUATION ?

- **VOUS TRAVAILLEZ**
  y compris vous êtes en congé de maladie ou de **maternité**, vous aidez un membre de votre famille dans son travail, vous êtes apprenti sous contrat, **stagiaire** rémunéré ☐ 1
  PASSEZ AU VERSO (questions 15 à 26)

- **VOUS NE TRAVAILLEZ PAS (OU PLUS),** répondez aux questions 12 à 14 et éventuellement 25 et 26 ☐ 2

(12) ÊTES-VOUS ?

- **Élève, étudiant, stagiaire non rémunéré** ☐ 1
- **Chômeur** *(inscrit ou non à l'ANPE)* ☐ 2
- **Retraité** *(ancien salarié)* ☐ 3
- **Préretraité** ☐ 4
- **Retiré des affaires** *(ancien agriculteur, ancien artisan, ancien commerçant, etc.)* ☐ 5
- **Femme** *(ou homme)* **au foyer,** *personne ne percevant qu'une pension de réversion, autre situation* ☐ 6

(13) CHERCHEZ-VOUS UN EMPLOI ?

- Vous ne cherchez pas d'emploi ☐ 1
- Vous cherchez un emploi depuis :
  - moins d'un an ☐ 2
  - plus d'un an ☐ 3

(14) AVEZ-VOUS DÉJÀ TRAVAILLÉ ?

OUI ☐ 1
NON ☐ 2

Si oui, quelle était votre profession principale ?

_____

Répondez aussi aux questions 25 et 26.-

Signez le bulletin au bas de la page 2 PAGE 1

VOUS TRAVAILLEZ

15 PROFESSION EXERCÉE ACTUELLEMENT

Soyez très précis. Exemples : électricien d'entretien de robot, comptable d'assurances niveau III, secrétaire de direction, technicien chimiste groupe IV coef. 225, etc.

16 TRAVAILLEZ-VOUS ?

A temps complet ☐ 1

A temps partiel :

Le temps partiel est déterminé par rapport au temps de travail normal dans votre entreprise

- plus d'un mi-temps ☐ 2
- à mi-temps ou moins ☐ 3

17 Si vous êtes agent de l'État, d'une collectivité locale, d'un hôpital public, d'un service public (EDF, SNCF, France Telecom, etc.) ou militaire de carrière, **PRÉCISEZ VOTRE CLASSIFICATION** *(corps, grade, etc.)* :

18 NOM (OU RAISON SOCIALE) DE L'ÉTABLISSEMENT QUI VOUS EMPLOIE OU QUE VOUS DIRIGEZ

Exemples : école Jules Ferry, cabinet du Docteur Moreau, Au Bonheur des Dames, Entreprise centrale du bâtiment, etc.

- Adresse de cet établissement :
  Exemple : 2, rue Hoche

Commune : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

Département : ☐☐ _____
(ou pays)

- Activité de cet établissement :
  ***Soyez très précis.** (Exemples : commerce en gros de fruits et légumes, fabrication d'outillage mécanique)*

19 ADRESSE DE VOTRE LIEU DE TRAVAIL **(si elle est différente de celle déclarée à la question 18)**

(Exemple : 18, boulevard Adolphe Pinard)
en cas de travail à domicile, indiquez « à domicile »

Commune : _____ Département : ☐☐
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

20 L'ÉLOIGNEMENT DE CE LIEU DE TRAVAIL **vous conduit-il à occuper régulièrement un deuxième logement (pied-à-terre, chambre d'hôtel, chez des parents...) ?**

OUI ☐ 1
NON ☐ 2

21 QUEL MOYEN DE TRANSPORT UTILISEZ-VOUS LE PLUS SOUVENT POUR ALLER TRAVAILLER ?

- Pas de transport *(travail à domicile)* ☐ 1
- Marche à pied seule ☐ 2
- Un seul moyen de transport
  - Voiture particulière ou deux-roues ☐ 3
  - Transports en commun ☐ 4
  - Plusieurs moyens de transports ☐ 5

22 VOTRE ACTIVITÉ PRINCIPALE CONSISTE-T-ELLE À AIDER UN MEMBRE DE VOTRE FAMILLE DANS SON TRAVAIL ? *(que vous perceviez ou non un salaire)*

(Exploitation agricole ou artisanale, commerce, profession libérale, etc.)

OUI ☐ 1
NON ☐ 2

23 ÊTES-VOUS ?

- Salarié *(y compris si vous aidez un membre de votre famille en étant rémunéré)* ☐ 1
- À votre compte *(y compris si vous aidez un membre de votre famille sans être rémunéré)* ☐ 2
- Chef de votre entreprise salarié, gérant majoritaire de SARL, gérant libre ☐ 3

LES QUESTIONS 24 À 26 NE S'ADRESSENT QU'AUX SALARIÉS ET AUX STAGIAIRES RÉMUNÉRÉS

24 INDIQUEZ VOTRE TYPE DE CONTRAT OU D'EMPLOI

- Apprenti sous contrat ☐ 1
- Placé par une agence d'intérim ☐ 2
- Contrat Emploi Solidarité ☐ 3
- Stagiaire rémunéré (SIFE, ...) ☐ 4
- Contrat de travail à durée déterminée (y compris contrat court, saisonnier...) ☐ 5
- Contrat à durée indéterminée, titulaire de la Fonction publique ☐ 6

25 INDIQUEZ LA POSITION PROFESSIONNELLE DE VOTRE EMPLOI

- Manœuvre, ouvrier spécialisé (OS1, OS2, OS3, etc.) ☐ 1
- Ouvrier qualifié ou hautement qualifié (P1, P2, P3, TA, OQ, etc.) ☐ 2
- Technicien, maîtrise administrative ou commerciale, VRP, dessinateur, programmeur, pupitreur ☐ 3
- Agent de maîtrise dirigeant des ouvriers ☐ 4
- Agent de maîtrise dirigeant des techniciens ou d'autres agents de maîtrise ☐ 5
- Instituteur, PEGC, assistant(e) social(e), infirmier(e) ☐ 6
- Ingénieur ou cadre d'entreprise *y compris nationalisée (les techniciens, agents de maîtrise ne devront pas se classer ici, même s'ils cotisent à une caisse de retraite des cadres)* ☐ 7
- Personnel de catégorie A de la Fonction publique ☐ 8
- Personnel de catégorie B de la Fonction publique ☐ 9
- Employé de bureau, employé de commerce, agent de service, aide soignant(e), personnel de catégorie C ou D de la Fonction publique ☐ 0

26 INDIQUEZ LA FONCTION PRINCIPALE DE VOTRE EMPLOI

- Production, fabrication, chantiers, exploitation ☐ 1
- Installation, entretien, réglage, réparation ☐ 2
- Nettoyage, gardiennage, travail ménager ☐ 3
- Manutention, magasinage, transports, logistique ☐ 4
- Secrétariat, saisie, guichet, standard, accueil ☐ 5
- Gestion, comptabilité, fonction administrative, organisation, RH, directeur général, état-major ☐ 6
- Commerce, vente, technico-commercial ☐ 7
- Recherche, études, méthodes, informatique ☐ 8
- Enseignement, santé, travail social, formation, documentation ☐ 9
- Information, publicité, arts, spectacles, sports ☐ 0

À _____, le _____ 1994.

Signature du déclarant :

3 Salles d'eau : (question 9 de la FL) — Indiquez pour chaque pièce utilisée pour la toilette l'existence de lavabo, baignoire, douche

Pièces utilisées pour la toilette	Existence			Si usage unique : toilette, cochez la case
	lavabo	douche	baignoire	Sinon : précisez (WC, cuisine ...)
1	☐	☐	☐	☐ _____
2	☐	☐	☐	☐ _____
3	☐	☐	☐	☐ _____
4	☐	☐	☐	☐ _____

4 Garage-Box-Parking (question 13 de la FL)

- Emplacement réservé en sous-sol ou garage attenant à la maison ou à l'immeuble ..... ☐ 1
- Cour privée reservée à l'immeuble avec réservation des places ..... ☐ 2
- Cour privée réservée à l'immeuble mais sans réservation des places ..... ☐ 3
- Emplacement réservé dépendant d' un immeuble (eu d'une propriété) voisin(e) ..... ☐ 4
- Pas d'emplacement réservé ..... ☐ 5

5 Des questions de la FL ou du BI ont-elles paru mal posées, obscures ?

oui ☐ 1 → lesquelles : _____
non ☐ 2 _____
_____

6 Le ménage a-t-il éprouvé des difficultés à donner une réponse à certaines questions :

oui ☐ 1 → lesquelles : _____
non ☐ 2 _____
_____

7 SI certaines questions sont non renseignées,

lesquelles : _____ pourquoi : _____

QUESTIONNAIRE COMPLEMENTAIRE

RECENSEMENT DE LA POPULATION
1994
APPLICATION PILOTE
A

Département : _____
Commune : _____

Interview
acceptée ☐ 1
refusée ☐ 2

DISTRICT : |__|__|__|__| |__|
Rang de l'immeuble : |__|__|__|
Rang du logement : |__|__|

Nom de l'occupant _____
Adresse _____

L'agent recenseur est-il intervenu (à la demande du ménage) dans le remplissage des questionnaires BI ou FL ? ☐ oui ☐ non

1 Surface du logement : (question 4 de la FL) — demander au ménage comment il est arrivé à la réponse qu'il a donnée, à reproduire ici _____ m²

- La ménage dispose d'un plan donnant la surface ..... ☐ 1
- Le ménage est locataire, l'information figure sur le bail ..... ☐ 2
- Le ménage dispose d'un autre document donnant la surface ..... ☐ 3
- Le ménage a mesuré la surface du logement pour l'occasion ..... ☐ 4
- Une estimation rapide de la surface de chaque pièce a été faite ..... ☐ 5
- L'estimation a été globale et très sommaire ..... ☐ 6
- L'un des membres du ménage "se souvenait" d'avoir eu connaissance de l'information précise ..... ☐ 7
- Le ménage a interrogé des voisins, dont le logement est ressemblant ..... ☐ 8
- Autre méthode, à préciser _____ ☐ 9
- Le ménage a refusé de répondre ..... ☐ 0

La réponse sur la FL : paraît exacte ☐ 1 ; paraît fausse ☐ 2 → estimation corrigée : _____ m² ; nature de l'erreur : _____

2 Année d'emménagement : (question 6 de la FL) — demander au ménage comment il est arrivé à la réponse

- Le ménage est locataire, l'information figure sur le bail ..... ☐ 1
- Le ménage est propriétaire, l'information est sur l'acte d'acquisition ..... ☐ 2
- Une des personnes du ménage a toujours habité ici ..... ☐ 3
- L'année d'emménagement correspond à une date importante pour le ménage (naissance, mariage ...) ..... ☐ 4
- Autre méthode, à préciser _____ ☐ 5
- Le ménage a refusé de répondre ..... ☐ 6

La réponse sur la FL : paraît exacte ☐ 1 ; paraît fausse ☐ 2 → estimation corrigée : _____

3 Année d'emménagement : (question 5 de la FL) — demander au ménage comment il est arrivé à la réponse

- Le ménage est locataire, l'information figure sur le bail ☐ 1
- Le ménage est propriétaire, l'information est sur l'acte d'acquisition ☐ 2
- Une des personnes du ménage a toujours habité ici ☐ 3
- L'année d'emménagement correspond à une date importante pour le ménage (naissance, mariage ...) ☐ 4
- Autre méthode, *précisez* ☐ 5
- Le ménage a refusé de répondre ☐ 6

La réponse sur la FL : paraît exacte ☐ 1 ; paraît fausse ☐ 2 → estimation corrigée : _____

4 Salles d'eau : (question 9 de la FL) — indiquez pour chaque pièce utilisée pour la toilette l'existence de lavabo, baignoire, douche

Pièces utilisées pour la toilette	Existence			Si usage unique : toilette, cochez la case. Sinon : précisez (WC, cuisine ...)
	lavabo	douche	baignoire	
1	☐	☐	☐	☐ _____
2	☐	☐	☐	☐ _____
3	☐	☐	☐	☐ _____
4	☐	☐	☐	☐ _____

5 Si le logement dispose d'une climatisation (oui à la question 12 de la FL)

Il s'agit :

- D'une climatisation centrale ☐ 1
- D'une climatisation pour une ou certaines pièces par appareil fixe ☐ 2
- D'une climatisation pour une ou certaines pièces par appareil mobile ☐ 3
- Autres cas, précisez ☐ 4

6 Si le ménage dispose d'une résidence secondaire (oui ou "blanc" à la question 14 de la FL)

- En est-il propriétaire ☐ 1
- S'agit-il d'une multipropriété ☐ 2
- En est-il locataire à l'année ☐ 3
- En a-t-il la jouissance à l'année ☐ 4
- Autres cas, précisez ☐ 5

Entre avril 1993 et mars 1994,

Cette résidence a-t-elle été utilisée par un ou des membres du ménage :

- Chaque week-end ☐ 1
- Plusieurs week-ends (par exemple: week-ends prolongés, week-ends d'été...) ☐ 2
- Seulement pour des périodes longues (une semaine ou plus) ☐ 3
- Jamais ☐ 4

Finalement, combien de fois cette résidence secondaire a-t-elle servi au ménage : _____

7 Des questions de la FL ou du BI ont-elles paru mal posées, obscures ?

oui ☐ 1 → lesquelles : _____

non ☐ 2

8 Le ménage a-t-il éprouvé des difficultés à donner une réponse à certaines questions :

oui ☐ 1 → lesquelles : _____

non ☐ 2

9 Si certaines questions sont non renseignées,

lesquelles : _____ pourquoi : _____

FEUILLE DE LOGEMENT

RECENSEMENT DE LA POPULATION

IMPRIMÉ NUMÉRO **1B**

À remplir pour tout logement d'habitation occupé ou non

Cadre à remplir par l'agent recenseur

Numéro du district de recensement

Rang de l'immeuble dans le district

Rang du logement dans l'immeuble

APPLICATION PILOTE B

1994

Remplissez les quatre pages très lisiblement.

(1) **NOM ET PRÉNOM DE L'OCCUPANT :** _____

(2) **ADRESSE DU LOGEMENT :** N° _____ Rue (ou lieudit) : _____

_____

Commune : _____ Département : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

(3) **LOCALISATION DU LOGEMENT DANS L'IMMEUBLE** (SI CE DERNIER COMPREND DEUX LOGEMENTS OU PLUS)

a. Escalier : _____ b. Étage : _____

c. Situation sur le palier : _____

_____ Si la porte d'entrée du logement a un numéro, indiquez ce numéro : _____

(4) **CATÉGORIE DE LOGEMENT**

1 ☐ **Résidence principale** (logement ou pièce indépendante où le ménage demeure la plus grande partie de l'année). Une chambre louée par un étudiant sera sa résidence principale. → Remplir un bulletin individuel pour chaque personne inscrite en liste 1, page 2.

2 ☐ **Logement** (ou pièce indépendante) **utilisé occasionnellement.**

3 ☐ **Résidence secondaire** ou logement loué (ou à louer) pour les loisirs ou les vacances. → Pas de bulletin individuel. Ne pas remplir les pages 2 et 3 de la feuille de logement.

4 ☐ **Logement vacant** (sans occupant, disponible ou non pour la vente ou la location). → Pas de bulletin individuel. Ne pas remplir les pages 2 et 3 de la feuille de logement.

RÉCAPITULATION GÉNÉRALE

à remplir par l'agent recenseur

(5)

Nombre d'imprimés n° 2 recueillis *(liste 1 seulement)* ☐

Enquête statistique non obligatoire.
Questionnaire confidentiel destiné à l'INSEE.
La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés garantit aux individus un droit d'accès et de rectification pour les informations les concernant. Ce droit d'accès peut être exercé pendant le délai où ces informations sont gardées sous forme nominative auprès des directions régionales de l'INSEE.

CARACTÉRISTIQUES GÉNÉRALES DE L'IMMEUBLE ET DU LOGEMENT

(1) TYPE DE LOGEMENT

- Logement-foyer pour personne(s) âgée(s) ☐ 1
- Chambre d'hôtel ☐ 2
- Construction provisoire, habitation de fortune ☐ 3
- Pièce indépendante (ayant sa propre entrée) ☐ 4
- Maison individuelle (y compris avec local professionnel), ferme ☐ 5
- Logement dans un immeuble collectif ☐ 6
- Logement dans un immeuble à usage autre que d'habitation *(usine, atelier, immeuble de bureaux, magasin, école, collège, hôpital, mairie, gare, bureau de poste, stade, etc.)* ☐ 7

(2) ANNÉE D'ACHÈVEMENT DE LA CONSTRUCTION DE LA MAISON OU DE L'IMMEUBLE ou année de la réhabilitation si la construction a fait l'objet d'une réhabilitation importante

- Avant 1915 ☐ 1
- de 1915 à 1948 ☐ 2
- de 1949 à 1967 ☐ 3
- de 1968 à 1974 ☐ 4
- de 1975 à 1981 ☐ 5
- de 1982 à 1989 ☐ 6
- 1990 ou après ☐ 7
- Immeuble en cours de construction partiellement habité ☐ 8

(3) NOMBRE DE PIÈCES D'HABITATION ⇨ ☐

- **Comptez les pièces telles que** chambres à coucher, salle à manger, salle de séjour, quelle que soit leur surface.
- **Ne comptez la cuisine** que si sa surface est supérieure à 12 m².
- **Ne comptez pas les pièces telles que** couloir, salle de bains, W.-C., buanderie, etc.
- **Ne comptez pas les pièces à usage exclusivement professionnel** *(exemples : cabinet de médecin, atelier d'artisan, etc.).*

(4) ÊTES-VOUS ?

- **Propriétaire du logement** *(y compris les différentes formes d'accession à la propriété)* ☐ 1
- **Locataire ou sous-locataire d'un logement loué vide.** ☐ 2
- **Locataire ou sous-locataire d'un logement loué meublé ou d'une chambre d'hôtel** ☐ 3
- Autre *(logé gratuitement, logé dans un appartement dont on a gardé la jouissance, etc.)* ☐ 4

(5) EN QUELLE ANNÉE AVEZ-VOUS EMMÉNAGÉ DANS CE LOGEMENT ? ⇨ P09-74 ☐

(Si tous les occupants actuels du logement ne sont pas arrivés en même temps, indiquez la date d'emménagement du premier arrivé. Si cette personne a toujours vécu dans le logement, reportez sa date de naissance.)

(6) LE LOGEMENT APPARTIENT-IL À UN ORGANISME HLM ? (office, société, ou OPAC) OUI ☐ 1 NON ☐ 2

(7) AVEZ-VOUS UNE BAIGNOIRE OU UNE DOUCHE ? OUI ☐ 1 NON ☐ 2

(8) W.-C.

Sont-ils situés à l'intérieur du logement ? OUI ☐ 1 NON ☐ 2

(9) NOMBRE DE SALLES D'EAU, *pièces réservées à la toilette contenant au moins une douche ou une baignoire.*

- Aucune ☐ 0
- 1 ☐ 1
- 2 ou plus ☐ 2

(10) MOYEN DE CHAUFFAGE DU LOGEMENT

- **Chauffage central collectif** *(commun à la totalité ou à la plupart des logements de l'immeuble, y compris le chauffage urbain).* ☐ 1
- **Chauffage central individuel** avec une chaudière propre au logement ☐ 2
- **Chauffage « tout électrique »** à radiateurs muraux ☐ 3
- **Autres moyens de chauffage** *(poêle, cheminée, cuisinière, radiateurs mobiles, appareils à accumulation, etc.)* ☐ 4

(11) COMBUSTIBLE PRINCIPAL UTILISÉ POUR LE CHAUFFAGE (quel que soit le moyen de chauffage du logement)

Cochez une seule case.

- Chauffage urbain ☐ 1
- Gaz de ville ou de réseau ☐ 2
- Fioul (mazout) ☐ 3
- Électricité ☐ 4
- Gaz en bouteilles ou citerne ☐ 5
- Charbon ☐ 6
- Bois ☐ 7

(12) CLIMATISATION

Le logement dispose-t-il ?

- D'aucune climatisation ☐ 1
- D'une climatisation centrale ☐ 2
- D'autres moyens de climatisation (climatisateurs mobiles, ...) *Ne pas compter les ventilateurs.* ☐ 3

(13) NOMBRE DE VOITURES DONT DISPOSENT LES HABITANTS DU LOGEMENT

- Aucune ☐ 0
- 1 ☐ 1
- 2 ☐ 2
- 3 ou plus ☐ 3

(14) DISPOSEZ-VOUS D'UNE RÉSIDENCE SECONDAIRE ?

(logement utilisé pour les loisirs, les vacances ou à titre occasionnel, mais pas pour des raisons liées à votre travail) OUI ☐ 1 NON ☐ 2

Que vous en soyez propriétaire ou non.

(15) CE LOGEMENT EST-IL LE SIÈGE D'UNE EXPLOITATION AGRICOLE ? OUI ☐ 1 NON ☐ 2

Si oui :

ⓐ **Superficie agricole utilisée :**

_____ **hectares,** _____ **ares**

Ne tenez pas compte des bois, étangs, terrains à bâtir, parcs et jardins d'agrément, landes et friches improductives, bâtiments et cours.

ⓑ **Orientation des productions agricoles**

Cochez une seule case.

- Exploitation avec une production principale
  - Polyculture (cultures de terres labourables) ☐ 1
  - Maraîchage ou horticulture ☐ 2
  - Vigne ou arbres fruitiers ☐ 3
  - Élevage d'herbivores (bovins, ovins) ☐ 4
  - Élevage de granivores (porcs, volailles) ☐ 5
- Autres cas ☐ 6

PAGE 4

BULLETIN INDIVIDUEL

RECENSEMENT DE LA POPULATION **1994**

IMPRIMÉ NUMÉRO **2B**

Ecrivez très lisiblement — Signez au bas de la page 2.

Cadre à remplir par l'INSEE

APPLICATION PILOTE 8

(1) NOM, Prénoms
Écrivez le nom en capitales
(exemple : ALLARD, épouse MAURIN, Françoise).

(2) ADRESSE

(3) SEXE
Masculin ☐ 1
Féminin ☐ 2

(4) ÉTAT MATRIMONIAL LÉGAL
Une personne vivant en couple non marié cochera la case correspondant à sa situation juridique présente : si, par exemple, elle est célibataire, elle cochera la 1re case.
Une personne en instance de divorce, ou séparée de son conjoint légitime, cochera la 2e case.
Célibataire ☐ 1
Marié(e) ou remarié(e) ☐ 2
Veuf(ve) ☐ 3
Divorcé(e) ☐ 4

(5) DATE ET LIEU DE NAISSANCE
Né(e) le : _____
(jour, mois, année)
à (commune) : _____
Département : _____
(pays pour l'étranger, territoire pour les TOM)

(6) NATIONALITÉ
- **Français de naissance** *(y compris par réintégration)* ☐ 1
- **Devenu français** ☐ 2
  Indiquez votre **nationalité antérieure :** _____
- **Étranger** ☐ 3
  Indiquez votre **nationalité :** _____

(7) OÙ HABITIEZ-VOUS LE 1er JANVIER 1987 ?
(pour toute personne née avant le 1er janvier 1987)
*Si, le 1er janvier 1987, vous étiez militaire ou élève interne, indiquez l'adresse de votre **résidence personnelle** à cette date et non celle de l'établissement (caserne, internat).*
- Dans le **même logement** que maintenant ☐ 1
- Dans un **autre logement** de la **même commune** ☐ 2
  (du même arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une **autre commune** ☐ 3
  (ou un autre arrondissement s'il s'agit de Paris, Lyon, Marseille)
  Indiquez cette **autre commune :**
  Commune : _____
  (pour Paris, Lyon, Marseille, précisez l'arrondissement)
  Département : _____
  (pays pour l'étranger, territoire pour les TOM)

LES QUESTIONS 8 À 26 NE S'ADRESSENT QU'AUX PERSONNES DE 14 ANS OU PLUS

(8) QUEL NIVEAU D'ÉTUDES AVEZ-VOUS ATTEINT ?
- Aucune scolarité ☐ 1
- Enseignement primaire ☐ 2
- Collège, classes de 6e à 3e, préparation CAP, BEP ☐ 3
- Classe de seconde, première ou terminale ☐ 4
- Enseignement supérieur ☐ 5

(9) INDIQUEZ TOUS LES DIPLÔMES QUE VOUS POSSÉDEZ
- Certificat d'études primaires ☐ 1
- BEPC, brevet élémentaire, brevet des collèges ☐ 2
- CAP ☐ 3
- BEP ☐ 4
- Baccalauréat général (Philo., Sciences Ex., Math. élém., A, B, C, D, E, L, S, ES...) ☐ 5
- Baccalauréat technologique (F, G, H, STI...) ou professionnel, brevet professionnel ou de technicien, autre brevet (BEA, BEC, BEI, etc.), capacité en droit ☐ 6
- Diplôme universitaire de 1er cycle, BTS, DEST, DUT, diplôme des professions sociales ou de la santé ☐ 7
- Diplôme universitaire de 2e ou 3e cycle, diplôme d'ingénieur, d'une grande école, etc. ☐ 8

(10) ÊTES-VOUS INSCRIT pour l'année scolaire 1993-94 dans un établissement d'enseignement supérieur ?
(y compris classes post-baccalauréat des lycées)
OUI ☐ 1
NON ☐ 2

(11) QUELLE EST VOTRE SITUATION ?
- **VOUS TRAVAILLEZ**
  y compris vous êtes en congé de maladie ou de maternité, vous aidez un membre de votre famille dans son travail, vous êtes apprenti sous contrat, stagiaire rémunéré ☐ 1
  PASSEZ AU VERSO (questions 15 à 26)
- **VOUS NE TRAVAILLEZ PAS (OU PLUS)**
  répondez aux questions 12 à 14 ☐ 2

(12) ÊTES-VOUS ?
- **Élève, étudiant, stagiaire non rémunéré** ☐ 1
  Précisez le lieu d'études ou de stage
  Commune : _____ *Département :* ☐☐
- **Chômeur** *(inscrit ou non à l'ANPE)* ☐ 2
- **Retraité** *(ancien salarié)* ☐ 3
- **Préretraité** *ou dispensé d'activité* ☐ 4
- **Retiré des affaires** *(ancien agriculteur, ancien artisan, ancien commerçant, etc.)* ☐ 5
- **Autre** *(femme ou homme au foyer, personne ne percevant qu'une pension de réversion, etc.)* ☐ 6

(13) CHERCHEZ-VOUS UN EMPLOI ?
- Vous ne cherchez pas d'emploi ☐ 1
- Vous cherchez un emploi depuis :
  - moins d'un an ☐ 2
  - plus d'un an ☐ 3

(14) AVEZ-VOUS DÉJÀ TRAVAILLÉ ?
OUI ☐ 1
NON ☐ 2
Si oui, quelle était votre profession principale ?
_____

Signez le bulletin au bas de la page 2 PAGE 1

VOUS TRAVAILLEZ

(15) PROFESSION EXERCÉE ACTUELLEMENT

Soyez très précis. *(Exemples : électricien* ***d'entretien de robot,*** *comptable* ***d'assurances niveau III,*** *secrétaire* ***de direction,*** *technicien* ***chimiste groupe IV coef. 225,*** *etc.)*

_____

_____

(16) TRAVAILLEZ-VOUS ?

À temps partiel ..... ☐ 1

À temps complet ..... ☐ 2

(17) VOTRE ACTIVITÉ PRINCIPALE CONSISTE-T-ELLE À AIDER UN MEMBRE DE VOTRE FAMILLE DANS SON TRAVAIL ? ***(que vous perceviez ou non un salaire)***

(Exploitation agricole ou artisanale, commerce, profession libérale, etc.)

OUI ☐ 1

NON ☐ 2

(18) Si vous êtes agent de l'État, d'une collectivité locale, d'un hôpital public, d'un service public (EDF, SNCF, France Telecom, etc.) ou militaire de carrière, PRÉCISEZ VOTRE CLASSIFICATION *(corps, grade, etc.)*

_____

_____

(19) ADRESSE DE VOTRE LIEU DE TRAVAIL

(Exemple : 18, boulevard Adolphe Pinard)

En cas de travail à domicile, indiquez « à domicile »

_____

Commune : _____

(pour Paris, Lyon, Marseille, précisez l'arrondissement)

Département : ☐☐ _____

(ou pays)

(20) NOM (OU RAISON SOCIALE) DE L'ÉTABLISSEMENT QUI VOUS EMPLOIE OU QUE VOUS DIRIGEZ

(Exemples : école Jules Ferry, cabinet du Docteur Moreau, Au Bonheur des Dames, Entreprise centrale du bâtiment, etc.)

En cas de travail chez un particulier, indiquez « particulier »

_____

_____

- **Activité de cet établissement :**

  Soyez très précis. *(Exemples : commerce* ***en gros*** *de fruits et légumes, fabrication d'outillage* ***mécanique,*** *etc.)*

_____

_____

- **Adresse de cet établissement** ***(si elle est différente de celle déclarée à la question 19)***

  (Exemple : 2, rue Hache)

_____

_____

Commune : _____ Département : ☐☐

(pour Paris, Lyon, Marseille, précisez l'arrondissement)

(21) QUEL MOYEN DE TRANSPORT UTILISEZ-VOUS LE PLUS SOUVENT POUR ALLER TRAVAILLER ?

- **Pas de transport** *(travail à domicile)* ..... ☐ 1
- **Marche à pied seule** ..... ☐ 2
- **Un seul moyen de transport**
  - Voiture particulière ou deux-roues ..... ☐ 3
  - Transports en commun ..... ☐ 4
- **Plusieurs moyens de transports**
  - Voiture et transports en commun ..... ☐ 5
  - Autres. *Précisez :* _____ ☐ 6

(22) ÊTES-VOUS ?

Salarié, y compris aide familial salarié ..... ☒ 1

À votre compte, y compris aide familial non salarié .... ☐ 2

- Gérant majoritaire de SARL, gérant libre, chef de votre propre entreprise salarié ..... ☐ 3

(23) SI VOUS ÊTES À VOTRE COMPTE

Combien de salariés employez-vous ?

Ne comptez ni les apprentis ni les gens de maison. Dans l'agriculture, comptez seulement les salariés permanents.

- Aucun ..... ☐ 1
- 1 ou 2 ..... ☐ 2
- 3 à 9 ..... ☐ 3
- 10 ou plus ..... ☐ 4

LES QUESTIONS 24 À 26 NE S'ADRESSENT QU'AUX SALARIÉS ET AUX STAGIAIRES RÉMUNÉRÉS

(24) INDIQUEZ VOTRE TYPE DE CONTRAT OU D'EMPLOI

- Apprenti sous contrat ..... ☐ 1
- Placé par une agence d'intérim ..... ☐ 2
- Contrat Emploi Solidarité ..... ☐ 3
- Stagiaire rémunéré (SIFE, ...) ..... ☐ 4
- Contrat de travail à durée déterminée (y compris contrat court, saisonnier...) ..... ☐ 5
- Contrat à durée indéterminée, titulaire de la Fonction publique ..... ☐ 6

(25) INDIQUEZ LA POSITION PROFESSIONNELLE DE VOTRE EMPLOI ACTUEL

- Manœuvre, ouvrier spécialisé (OS1, OS2, OS3, etc.) ... ☐ 1
- Ouvrier qualifié ou hautement qualifié (P1, P2, P3, TA, OQ, etc.) ..... ☐ 2
- Technicien, maîtrise administrative ou commerciale, VRP, dessinateur, programmeur, pupitreur ..... ☐ 3
- Agent de maîtrise dirigeant des ouvriers ..... ☐ 4
- Agent de maîtrise dirigeant des techniciens ou d'autres agents de maîtrise ..... ☐ 5
- Instituteur, PEGC, assistant(e) social(e), infirmier(e) et personnel de catégorie B de la Fonction publique ..... ☐ 6
- Ingénieur ou cadre d'entreprise y compris nationalisée *(les techniciens, agents de maîtrise ne devront pas se classer ici, même s'ils cotisent à une caisse de retraite des cadres)* ..... ☐ 7
- Professeur et personnel de catégorie A de la Fonction publique ..... ☐ 8
- Employé de bureau, employé de commerce, agent de service, aide soignant(e), personnel de catégorie C ou D de la Fonction publique ..... ☐ 9
- Autre cas. *Précisez :* _____ ☐ 0

(26) QUELLE EST VOTRE FONCTION PRINCIPALE ?

Cochez la case et entourez la fonction correspondant à votre emploi

- Production fabrication chantiers exploitation ..... ☐ 1
- Installation entretien réglage réparation ..... ☐ 2
- Nettoyage gardiennage travail ménager ..... ☐ 3
- Manutention magasinage transports logistique ..... ☐ 4
- Secrétariat saisie guichet standard accueil ..... ☐ 5
- Gestion comptabilité fonction administrative organisation RH état-major ..... ☐ 6
- Commerce vente technico-commercial ..... ☐ 7
- Recherche études méthodes informatique ..... ☐ 8
- Enseignement santé travail social formation documentation ..... ☐ 9
- Information publicité arts spectacles sports ..... ☐ 0

Enquête statistique non obligatoire.
Questionnaire confidentiel destiné à l'INSEE.
La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés garantit aux individus un droit d'accès et de rectification pour les informations les concernant. Ce droit d'accès peut être exercé pendant le délai où ces informations sont gardées sous forme nominative auprès des directions régionales de l'INSEE.

À _____ le _____ 1994.

Signature du déclarant :

	Nom Prénom					
1 Niveau d'études *Report de la réponse à la question 8 du BI* *Si la réponse est à "blanc", s'agit-il :* • D'un oubli • D'un refus • D'une incompréhension, *précisez pourquoi* *Si la réponse 8 est 1, expliquez*	☐☐ ☐ 1 ☐ 2 ☐ 3 _____	☐☐	☐☐	☐☐	☐☐	☐☐
2 Diplômes *Report de la réponse à la question 9 du BI* *Si la question 9 est à blanc, s'agit-il :* • D'un oubli • D'un refus • Absence de diplôme • Autre, *précisez pourquoi* *Dans tous les cas, reposer la question :* Avez-vous un ou plusieurs diplômes ? Lesquels ? (ce distinguent, formation continue...) *Chiffrement (diplôme le plus élevé)*	☐☐ ☐ 1 ☐ 2 ☐ 3 ☐ 4 _____ _____ ☐☐	☐☐	☐☐	☐☐	☐☐	☐☐
Si la personne travaille						
5 Quel(s) moyen(s) de transport la personne utilise-t-elle pour aller travailler ? *Chiffrement* *Entourer la bonne réponse*	1er 2ème 3ème 1 2 3 4 5 6	1er 2ème 3ème 1 2 3 4 5 6	1er 2ème 3ème 1 2 3 4 5 6	1er 2ème 3ème 1 2 3 4 5 6	1er 2ème 3ème 1 2 3 4 5 6	1er 2ème 3ème 1 2 3 4 5 6

QUESTIONNAIRE COMPLEMENTAIRE

RECENSEMENT DE LA POPULATION

1994

APPLICATION PILOTE

B

Département : _____

Commune : _____

Interview
acceptée ☐ 1
refusée ☐ 2

DISTRICT : |__|__|__|__|__|__|

Rang de l'immeuble : |__|__|__|

Rang du logement : |__|__|

Nom de l'occupant _____

Adresse _____

L'agent recenseur est-il intervenu (à la demande du ménage) dans le remplissage des questionnaires BI et FL ? ☐ oui ☐ non

1 Année d'achèvement de l'immeuble (question 2 de la FL)

L'immeuble a-t-il fait l'objet d'une réhabilitation importante ?..... oui ☐ 1
non ☐ 2
ne sait pas ☐ 3

Si oui :

- Cette réhabilitation a-t-elle été prise en compte dans la réponse ?..... oui ☐ 1
non ☐ 2
- Quelle a été la nature de la réhabilitation ?
  - Ajout d'un bâtiment..... ☐ 1
  - Modification du gros oeuvre..... ☐ 2
  - Création de nouveaux appartements..... ☐ 3
  - Autres, précisez _____ ☐ 4

2 Statut d'occupation (question 4 de la FL)

Si la case 4 est cochée, dans quelle situation le ménage se trouve-t-il ?

- Logé entièrement gratuitement..... ☐ 1
- Avantage en nature..... ☐ 2
- Logement vendu en viager..... ☐ 3
- Usufruitier..... ☐ 4
- Autre, précisez : ..... ☐ 5

Nom Prénom	
	I_I_I
1 Année d'entrée en France	
a) L'année n'est pas renseignée • Sans objet (né en France) • Oubli de répondre • Ne sait pas • Hésite entre plusieurs dates • Refus de répondre	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5
b) L'année est renseignée • La personne n'a pas hésité	☐ 1
• La personne a hésité entre plusieurs dates car elle a fait plusieurs voyages, et elle a finalement choisi : - la date la plus ancienne - la date la plus récente - une date intermédiaire	☐ 2 ☐ 3 ☐ 4
• Autres cas, précisez	☐ 5 _____
2 Niveau d'études et diplôme le plus élevé	9 I_I 10 I_I
Report des réponses aux questions 9 et 10 du BI	
Si la question 9 est "à blanc", s'agit-il : • D'un oubli • D'un refus • D'une absence totale de scolarisation, *précisez pourquoi* • D'une incompréhension, *précisez pourquoi*	☐ 1 ☐ 2 ☐ 3 _____ ☐ 4 _____
Si la question 10 est "à blanc", s'agit-il : • D'un oubli • D'un refus • Autre, *précisez pourquoi*	☐ 1 ☐ 2 ☐ 3 _____
Dans tous les cas, reposez la question : Avez-vous un ou plusieurs diplômes ? Lesquels ? (yc étrangers, formation continue...)	_____
Chiffrement (diplôme le plus élevé)	I_I
Si la personne travaille	
3 Temps de travail de la personne • Temps de travail hebdomadaire (en heures) de la personne • Temps de travail hebdomadaire normal dans l'entreprise ou la profession	_____ _____
4 Lieu de travail	oui ☐ 1 non ☐ 2
Report de la réponse à la question 20 du BI	
La personne effectue-t-elle tous les jours où elle travaille le trajet domicile-travail ?	oui ☐ 1 non ☐ 2
Si non :	
- Où la personne loge-t-elle quand elle ne rentre pas chez elle ? • A l'hôtel • Chez des parents ou amis • Dans un logement qu'elle loue ou qui lui appartient • Autre cas, précisez	☐ 1 ☐ 2 ☐ 3 ☐ 4 _____
- Combien de fois par semaine la personne ne loge-t-elle pas chez elle ?	
La question 20 a-t-elle été comprise ?	oui ☐ 1 non ☐ 2
5 Quel(s) moyen(s) de transport la personne utilise-t-elle pour aller travailler ?	1 à pied ☐ 2 vélo ☐ 3 moto ☐ 4 voiture ☐ 5 TC A1 ☐ ☐ ☐ ☐ ☐ A2 ☐ ☐ ☐ ☐ ☐ A3 ☐ ☐ ☐ ☐ ☐ A4 ☐ ☐ ☐ ☐ ☐ A5 ☐ ☐ ☐ ☐ ☐
Chiffrement *Entourer la bonne réponse*	1 2 3 4 5

I_I_I	I_I_I	I_I_I
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5	☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5
☐ 1	☐ 1	☐ 1
☐ 2 ☐ 3 ☐ 4 ☐ 5 _____	☐ 2 ☐ 3 ☐ 4 ☐ 5 _____	☐ 2 ☐ 3 ☐ 4 ☐ 5 _____
9 I_I 10 I_I	9 I_I 10 I_I	9 I_I 10 I_I
☐ 1 ☐ 2 ☐ 3 _____ ☐ 4 _____	☐ 1 ☐ 2 ☐ 3 _____ ☐ 4 _____	☐ 1 ☐ 2 ☐ 3 _____ ☐ 4 _____
☐ 1 ☐ 2 ☐ 3 _____	☐ 1 ☐ 2 ☐ 3 _____	☐ 1 ☐ 2 ☐ 3 _____
_____	_____	_____
I_I	I_I	I_I
_____ _____	_____ _____	_____ _____
oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2
oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2
☐ 1 ☐ 2 ☐ 3 ☐ 4 _____	☐ 1 ☐ 2 ☐ 3 ☐ 4 _____	☐ 1 ☐ 2 ☐ 3 ☐ 4 _____
oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2	oui ☐ 1 non ☐ 2
1 à pied ☐ 2 vélo ☐ 3 moto ☐ 4 voiture ☐ 5 TC A1 ☐ ☐ ☐ ☐ ☐ A2 ☐ ☐ ☐ ☐ ☐ A3 ☐ ☐ ☐ ☐ ☐ A4 ☐ ☐ ☐ ☐ ☐ A5 ☐ ☐ ☐ ☐ ☐ 1 2 3 4 5	1 à pied ☐ 2 vélo ☐ 3 moto ☐ 4 voiture ☐ 5 TC A1 ☐ ☐ ☐ ☐ ☐ A2 ☐ ☐ ☐ ☐ ☐ A3 ☐ ☐ ☐ ☐ ☐ A4 ☐ ☐ ☐ ☐ ☐ A5 ☐ ☐ ☐ ☐ ☐ 1 2 3 4 5	1 à pied ☐ 2 vélo ☐ 3 moto ☐ 4 voiture ☐ 5 TC A1 ☐ ☐ ☐ ☐ ☐ A2 ☐ ☐ ☐ ☐ ☐ A3 ☐ ☐ ☐ ☐ ☐ A4 ☐ ☐ ☐ ☐ ☐ A5 ☐ ☐ ☐ ☐ ☐ 1 2 3 4 5

FEUILLE DE LOGEMENT

RECENSEMENT DE LA POPULATION

1995

APPLICATION PILOTE

IMPRIMÉ NUMÉRO **1**

À remplir pour tout logement d'habitation, occupé ou non.

Cadre à remplir par l'agent recenseur

Numéro du district de recensement

Rang de l'immeuble dans le district

Rang du logement dans l'immeuble

Remplissez les quatre pages très lisiblement.

(1) **NOM ET PRÉNOM DE L'OCCUPANT :** _____

(2) **ADRESSE DU LOGEMENT :** N° _____ Rue (ou lieudit) : _____

_____

Commune : _____ Département : _____

(pour Paris, Lyon, Marseille, précisez l'arrondissement)

(3) **LOCALISATION DU LOGEMENT DANS L'IMMEUBLE** (SI CE DERNIER COMPREND DEUX LOGEMENTS OU PLUS)

a. Escalier : _____ *b.* Étage : _____

c. Situation sur le palier : _____

_____ Si la porte d'entrée du logement a un numéro, indiquez ce numéro : _____

(4) **CATÉGORIE DE LOGEMENT**

1 ☐ **Résidence principale** (logement ou pièce indépendante où le ménage demeure la plus grande partie de l'année). Une chambre louée par un étudiant sera sa résidence principale.

⇨ Remplir un bulletin individuel pour chaque personne inscrite en liste A, page 2.

2 ☐ **Logement occasionnel** (logement ou pièce indépendante utilisé comme habitation pour des raisons professionnelles par une personne ayant sa résidence principale ailleurs).

3 ☐ **Résidence secondaire** ou logement loué (ou à louer) pour les loisirs ou les vacances.

⇨ Pas de bulletin individuel sauf si des personnes habitent le logement au moment du recensement. Dans ce cas, établir les bulletins individuels correspondants, en prenant soin de remplir le cadre "PERSONNES DE PASSAGE".
Ne pas remplir les pages 2 et 3 de la feuille de logement.

4 ☐ **Logement vacant** (sans occupant, disponible ou non pour la vente ou la location).

⇨ Pas de bulletin individuel.
Ne pas remplir les pages 2 et 3 de la feuille de logement.

5 ☐ **Caravane, habitation mobile.**

⇨ Remplir un bulletin individuel pour chaque personne inscrite en liste A.
Ne pas remplir la page 4 ; à classer à part.

RÉCAPITULATION GÉNÉRALE

À remplir par l'agent recenseur	*À remplir par la mairie*			
(5) Nombre d'imprimés n° 2 recueillis *(liste A seulement)*	(6) Nombre d'imprimés n° 2 supprimés *(faisant double emploi avec des imprimés n° 5 réintégrés ou des domiciliations effectuées)*	(7) Nombre d'imprimés n° 2 ajoutés *(bulletins reçus pour des personnes en déplacement et ne faisant pas double emploi)*	(8) Nombre d'imprimés n° 5 réintégrés	(9) Population après réintégrations (9) = (5) − (6) + (7) + (8)

Nomtre d'imprimés n° 2 *bis* établis

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête, reconnue d'intérêt général, est obligatoire.
Visa n° 9SX056 Ec du ministre de l'Economie (Insee) valable pour l'année 1995.
Selon la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques, tout défaut de réponse ou une réponse sciemment inexacte peut entraîner l'appli[illegible] d'une amende administrative.
Questionnaire confidentiel destiné à l'Insee.
La loi n° 78-17 du 6 janvier 1978, relative à l'informatique, aux fichiers et aux libertés, garantit aux individus un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des Directions Régionales de l'Insee.

CARACTÉRISTIQUES GÉNÉRALES DE L'IMMEUBLE ET DU LOGEMENT

(1) TYPE DE LOGEMENT

- Logement-foyer pour personne(s) âgée(s) ☐ 1
- Chambre d'hôtel ☐ 2
- Construction provisoire, habitation de fortune ☐ 3
- Pièce indépendante (ayant sa propre entrée) ☐ 4
- Maison individuelle, ferme ☐ 5
- Logement dans un immeuble collectif ☐ 6
- Logement dans un immeuble à usage autre que d'habitation *(usine, atelier, immeuble de bureaux, magasin, école, collège, hôpital, mairie, gare, bureau de poste, stade, etc.)* ☐ 7

(2) ANNÉE D'ACHÈVEMENT DE LA CONSTRUCTION DE LA MAISON OU DE L'IMMEUBLE

- Avant 1915 ☐ 1
- de 1915 à 1948 ☐ 2
- de 1949 à 1967 ☐ 3
- de 1968 à 1974 ☐ 4
- de 1975 à 1981 ☐ 5
- de 1982 à 1989 ☐ 6
- 1990 ou après ☐ 7

Dans ce cas, précisez l'année d'achèvement ⇨ 19 ☐☐

- Immeuble en cours de construction partiellement habité ☐ 8

Si les différentes parties ne sont pas de la même époque, indiquez l'année d'achèvement de la partie habitée ou de la partie habitée la plus importante.

(3) NOMBRE DE PIÈCES D'HABITATION ⇨ ☐☐

- **Comptez les pièces telles que** salle à manger, séjour, chambres, quelle que soit leur surface.
- **Ne comptez la cuisine** que si sa surface est supérieure à 12 m².
- **Ne comptez pas les pièces telles** que couloir, salle de bains, W.-C., buanderie, etc.
- **Ne comptez pas les pièces à usage exclusivement professionnel** *(exemples : cabinet de médecin, atelier d'artisan, etc.).*

(4) SURFACE DU LOGEMENT

- **Tenez compte de toutes les pièces,** y compris couloir, cuisine, salle de bains, W.-C., etc.
- **Ne tenez pas compte des balcons, terrasses, vérandas, caves, parkings, greniers.**
- moins de 40 m² ☐ 1
- de 40 à moins de 70 m² ☐ 2
- de 70 à moins de 100 m² ☐ 3
- de 100 à moins de 150 m² ☐ 4
- 150 m² ou plus ☐ 5

(5) ÊTES-VOUS ?

- **Propriétaire du logement** *(y compris les différentes formes d'accession à la propriété)* ☐ 1
- **Locataire ou sous-locataire d'un logement loué vide** ☐ 2
- **Locataire ou sous-locataire d'un logement loué meublé ou d'une chambre d'hôtel** ☐ 3
- **Logé gratuitement,** par exemple par des parents, des amis ou votre employeur *(y compris le cas des personnes occupant un logement qu'elles ont vendu en viager ou dont elles ont la jouissance par usufruit)* ☐ 4

(6) EN QUELLE ANNÉE AVEZ-VOUS EMMÉNAGÉ DANS CE LOGEMENT ? ⇨ 19 ☐☐

(Si tous les occupants actuels du logement ne sont pas arrivés en même temps, indiquez la date d'emménagement du premier arrivé. Si cette personne a toujours vécu dans le logement, reportez sa date de naissance.)

(7) LE LOGEMENT APPARTIENT-IL À UN ORGANISME HLM ? (office, société ou OPAC)

OUI ☐ 1
NON ☐ 2

(8) INSTALLATIONS SANITAIRES

- **Ni baignoire, ni douche** dans le logement ☐ 0
- Baignoire ou douche dans une **pièce non réservée** à la toilette ☐ 1
- **1 salle d'eau** (pièce réservée à la toilette contenant au moins baignoire ou douche) ☐ 2
- **2 salles d'eau** ou plus ☐ 3

(9) W.-C.

Sont-ils situés à l'intérieur du logement ?

OUI ☐ 1
NON ☐ 2

(10) ÉVACUATION DES EAUX USÉES

- Tout à l'égout ☐ 1
- Fosse septique *(pour W.-C. et eaux ménagères)* ☐ 2
- Autres cas ☐ 3

(11) MOYEN DE CHAUFFAGE DU LOGEMENT

- **Chauffage central collectif** *(commun à la totalité ou à la plupart des logements de l'immeuble, y compris le chauffage urbain)* ☐ 1
- **Chauffage central individuel** avec une chaudière propre au logement ☐ 2
- **Chauffage «tout électrique»** à radiateurs muraux ☐ 3
- **Autres moyens de chauffage** *(poêle, cheminée, cuisinière, radiateurs mobiles, appareils à accumulation, etc.)* ☐ 4

(12) COMBUSTIBLE PRINCIPAL UTILISÉ POUR LE CHAUFFAGE *(Cochez une seule case).*

- Chauffage urbain ☐ 1
- Gaz de ville ou de réseau ☐ 2
- Fioul (mazout) ☐ 3
- Électricité ☐ 4
- Gaz en bouteilles ou citerne ☐ 5
- Charbon ☐ 6
- Bois ☐ 7

(13) GARAGE - BOX - PARKING

Disposez-vous, pour votre usage personnel, d'un **emplacement réservé de stationnement** situé dans l'immeuble ou la propriété ?

OUI ☐ 1
NON ☐ 2

(14) NOMBRE DE VOITURES DONT DISPOSENT LES HABITANTS DU LOGEMENT

- Aucune ☐ 0
- 1 ☐ 1
- 2 ou plus ☐ 2

(15) LE LOGEMENT EST-IL LE SIÈGE D'UNE EXPLOITATION AGRICOLE ?

OUI ☐ 1
NON ☐ 2

Si oui :

(a) Superficie agricole utilisée :

_____ hectares _____ ares

Ne tenez pas compte des bois, étangs, terrains à bâtir, parcs et jardins d'agrément, landes et friches improductives, bâtiments et cours.

(b) Orientation des productions agricoles

- Exploitation avec une production principale
  - Polyculture (cultures de terres labourables) ☐ 1
  - Maraîchage ou horticulture ☐ 2
  - Vigne ou arbres fruitiers ☐ 3
  - Élevage d'herbivores (bovins, ovins) ☐ 4
  - Élevage de granivores (porcs, volailles) ☐ 5
- Autres cas
  - Polyculture - Élevage ☐ 6
  - Élevage d'herbivores et de granivores ☐ 7
  - Autres ☐ 8

BULLETIN INDIVIDUEL

RECENSEMENT DE LA POPULATION

1995

APPLICATION PILOTE

IMPRIMÉ NUMÉRO **2**

Entourez le numéro correspondant à votre situation

Signez au bas de la page 2

1 NOM, Prénoms
Écrivez le nom en capitales
(exemple : ALLARD, épouse MAURIN, Françoise).

2 ADRESSE

3 SEXE
Masculin ☐ 1
Féminin ☐ 2

4 ÉTAT MATRIMONIAL LÉGAL
Une personne vivant en union libre cochera la case correspondant à sa situation juridique présente : si, par exemple, elle est célibataire, elle cochera la 1re case.
Une personne en instance de divorce, ou séparée de son conjoint légitime, cochera la 2e case.
Célibataire ☐ 1
Marié(e) ou remarié(e) ☐ 2
Veuf(ve) ☐ 3
Divorcé(e) ☐ 4

5 DATE ET LIEU DE NAISSANCE
Né(e) le : _____
(jour, mois, année)
à (commune) : _____
Département : _____
(pays pour l'étranger, territoire pour les TOM)
Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ? _____

6 NATIONALITÉ
- Français de naissance (y compris par réintégration) ☐ 1
- Devenu français ☐ 2
  Indiquez votre nationalité antérieure : _____
- Étranger ☐ 3
  Indiquez votre nationalité : _____

7 ÊTES-VOUS INSCRIT(E) pour l'année scolaire 1995-1996 dans un établissement d'enseignement ?
OUI ☐ 1
NON ☐ 2
Si oui, précisez le lieu d'études :
Commune : _____ Département : ☐☐

8 OÙ HABITIEZ-VOUS LE 1er JANVIER 1986 ?
(pour toute personne née avant le 1er janvier 1986)
Si, le 1er janvier 1986, vous étiez militaire ou élève interne, indiquez l'adresse de votre résidence personnelle à cette date et non celle de l'établissement (caserne, internat).
- Dans le même logement que maintenant ☐ 1
- Dans un autre logement de la même commune ☐ 2
  (du même arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une autre commune ☐ 3
  (ou un autre arrondissement s'il s'agit de Paris, Lyon, Marseille)

Indiquez cette autre commune :
Commune : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)
Département : _____
(pays pour l'étranger, territoire pour les TOM)

LES QUESTIONS 9 À 25 NE S'ADRESSENT QU'AUX PERSONNES DE 14 ANS OU PLUS

9 QUEL NIVEAU D'ÉTUDES AVEZ-VOUS ATTEINT ?
- École primaire ☐ 1
- Collège, classes de 6e à 3e, CAP, BEP ☐ 2
- Classes de seconde, première ou terminale ☐ 3
- Études supérieures (facultés, IUT, etc.) ☐ 4

10 INDIQUEZ VOTRE DERNIER DIPLÔME OBTENU
- Aucun diplôme ☐ 0
- Certificat d'études primaires ☐ 1
- BEPC, brevet élémentaire, brevet des collèges ☐ 2
- CAP ☐ 3
- BEP ☐ 4
- Baccalauréat général (Philo., Sciences ex., Math. élém., A, B, C, D, E, L, S, ES, etc.) ☐ 5
- Baccalauréat technologique (F, G, H, STI, etc.) ou professionnel, brevet professionnel ou de technicien, autre brevet (BEA, BEC, BEI, etc.), capacité en droit ☐ 6
- Diplôme universitaire de 1er cycle, BTS, DUT, diplôme des professions sociales ou de la santé ☐ 7
- Diplôme universitaire de 2e ou 3e cycle, diplôme d'ingénieur, d'une grande école, etc. ☐ 8

11 QUELLE EST VOTRE SITUATION ?
- ***VOUS TRAVAILLEZ***
  Cochez la case et passez au verso (questions 15 à 25)
  y compris si vous êtes en congé de maladie ou de maternité, si vous aidez un membre de votre famille dans son travail ou si vous êtes apprenti sous contrat, stagiaire rémunéré ☐ 1
- ***VOUS NE TRAVAILLEZ PAS (OU PLUS)***
  Cochez la case et répondez aux questions 12 à 14 ☐ 2

12 ÊTES-VOUS ?
- Étudiant *(facultés, IUT, grandes écoles, etc.)* ☐ 1
- Élève *(collège, lycée)* ☐ 2
- Stagiaire non rémunéré ☐ 3
- Chômeur *(inscrit ou non à l'ANPE)* ☐ 4
- Préretraité ☐ 5
- Retraité *(ancien salarié)* ☐ 6
- Retiré des affaires *(ancien agriculteur, ancien artisan, ancien commerçant, etc.)* ☐ 7
- Autre *(femme ou homme au foyer, personne ne percevant qu'une pension de réversion ou d'invalidité, etc.)* ☐ 8

13 CHERCHEZ-VOUS UN EMPLOI ?
- Vous ne cherchez pas d'emploi ☐ 1
- Vous cherchez un emploi depuis :
  - moins d'un an ☐ 2
  - plus d'un an ☐ 3

14 AVEZ-VOUS DÉJÀ TRAVAILLÉ ?
OUI ☐ 1 → Quelle était votre profession principale ? _____
NON ☐ 2

Signez le bulletin au bas de la page 2

POUR LES PERSONNES DE PASSAGE (voir page 3 de l'imprimé n° 1), adresse de la résidence habituelle :
N° _____ Rue (ou lieudit) : _____ Commune : 39_____ Département : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

PAGE 1

VOUS TRAVAILLEZ

15 PROFESSION EXERCÉE ACTUELLEMENT
Soyez très précis. Exemples : électricien d'entretien de robot, comptable d'assurances niveau III, secrétaire de direction, technicien chimiste groupe IV coef. 225, etc.

16 TRAVAILLEZ-VOUS ?
Le temps partiel est déterminé par rapport au temps de travail normal dans votre entreprise

- À temps complet ..... ☐ 1
- À temps partiel :
  - plus d'un mi-temps ... ☐ 2
  - à mi-temps ou moins . ☐ 3

17 VOTRE ACTIVITÉ PRINCIPALE CONSISTE-T-ELLE À AIDER UN MEMBRE DE VOTRE FAMILLE DANS SON TRAVAIL ? *(que vous perceviez ou non un salaire)*
(Exploitation agricole ou artisanale, commerce, profession libérale, etc.)
OUI ☐ 1
NON ☐ 2

18 Si vous êtes agent de l'État, d'une collectivité locale, d'un hôpital public, d'un service public (EDF, SNCF, La Poste, etc.) **ou militaire de carrière, PRÉCISEZ VOTRE CLASSIFICATION** *(corps, grade, etc.)*

19 OÙ TRAVAILLEZ-VOUS ?

ⓐ **Adresse de votre lieu de travail :**
(Exemple : 18, boulevard Adolphe Pinard)
En cas de travail à domicile, indiquez « à domicile »
En cas de travail chez un particulier, indiquez « particulier »

Commune : _____
(pour Paris, Lyon, Marseille, précisez l'arrondissement)

Département : ☐☐ _____
(ou pays)

ⓑ **Nom (ou raison sociale) de l'établissement qui vous emploie ou que vous dirigez :**

ⓒ **Adresse de cet établissement**, si elle est différente de celle déclarée à la question 19 a.

ⓓ **Activité de cet établissement :**
Soyez très précis. Exemples : commerce en gros de fruits et légumes, fabrication d'outillage mécanique, etc.

20 QUEL MODE DE TRANSPORT UTILISEZ-VOUS LE PLUS SOUVENT POUR ALLER TRAVAILLER ?

- Pas de transport *(travail à domicile)* ☐ 1
- Marche à pied uniquement ☐ 2
- Un seul mode de transport
  - Deux-roues ☐ 3
  - Voiture particulière ☐ 4
  - Transports en commun ☐ 5
- Plusieurs modes de transport ☐ 6

21 ÊTES-VOUS ?

- Indépendant ou à votre compte, y compris aide familial non salarié ☐ 1
- Chef d'entreprise salarié, PDG, gérant minoritaire de SARL, co-gérant ☐ 2
- Salarié, y compris aide familial salarié ☐ 3

22 SI VOUS ÊTES À VOTRE COMPTE OU CHEF D'ENTREPRISE
Combien de salariés employez-vous ?
Ne comptez ni les apprentis ni les gens de maison. Dans l'agriculture, comptez seulement les salariés permanents.

- Aucun ... ☐ 1
- 1 ou 2 ... ☐ 2
- 3 à 9 ..... ☐ 3
- 10 ou plus ☐ 4

LES QUESTIONS 23 À 25 NE S'ADRESSENT QU'AUX SALARIÉS ET AUX STAGIAIRES RÉMUNÉRÉS

23 INDIQUEZ VOTRE TYPE DE CONTRAT OU D'EMPLOI

- Apprenti sous contrat ☐ 1
- Placé par une agence d'intérim ☐ 2
- Contrat Emploi Solidarité (CES) ☐ 3
- Stagiaire rémunéré (SIFE, etc.) ☐ 4
- Contrat de travail à durée déterminée (y compris contrat court, saisonnier, etc.) ☐ 5
- Emploi à durée indéterminée ☐ 6
- Titulaire de la Fonction publique ☐ 7

24 INDIQUEZ LA CATÉGORIE PROFESSIONNELLE DE VOTRE EMPLOI

- Manœuvre, ouvrier spécialisé (OS1, OS2, OS3, etc.) . ☐ 1
- Ouvrier qualifié ou hautement qualifié (P1, P2, P3, OQ, OHQ, TA, etc.) ☐ 2
- Agent de service, aide soignant(e), femme de ménage, employé de commerce ☐ 3
- Employé de bureau, agent de catégorie C ou D de la Fonction publique ☐ 4
- Agent de maîtrise, contremaître dirigeant des ouvriers . ☐ 5
- Agent de maîtrise dirigeant des techniciens ou d'autres agents de maîtrise (y compris cotisant à une retraite de cadres) ☐ 6
- Technicien, dessinateur, programmeur, pupitreur, maîtrise administrative, comptable, commerciale d'entreprise, VRP ☐ 7
- Instituteur, assistant(e) social(e), infirmier(e), technicien médical, agent de catégorie B de la Fonction publique . ☐ 8
- Cadre ou ingénieur d'entreprise, professeur, agent de catégorie A ou assimilé de la Fonction publique . . ☐ 9
- Autre cas. *Précisez :* _____ ☐ 0

25 INDIQUEZ LA FONCTION PRINCIPALE DE VOTRE EMPLOI

- Production, fabrication, chantiers, exploitation ☐ 1
- Installation, entretien, réglage, réparation ☐ 2
- Nettoyage, gardiennage, travail ménager ☐ 3
- Manutention, magasinage, transports, logistique .... ☐ 4
- Secrétariat, saisie, guichet, standard, accueil ☐ 5
- Gestion, comptabilité, fonction administrative, organisation, personnel, directeur général, état-major ☐ 6
- Commerce, vente, technico-commercial ☐ 7
- Recherche, études, méthodes, informatique ☐ 8
- Enseignement, santé, travail social, formation, documentation ☐ 9
- Information, publicité, arts, spectacles, sports ☐ 0

Nous vous remercions de votre participation

Vu l'avis favorable du Conseil National de l'Information Statistique, cette enquête, reconnue d'intérêt général, est obligatoire.
Visa n° 95X056 Ec du ministre de l'Économie (Insee) valable pour l'année 1995.
Selon la loi n° 51-711 du 7 juin 1951 modifiée sur l'obligation, la coordination et le secret en matière de statistiques, tout défaut de réponse ou une réponse sciemment inexacte peut entraîner l'application d'une amende administrative.
Questionnaire confidentiel destiné à l'Insee.
La loi n° 78-17 du 6 janvier 1978, relative à l'informatique, aux fichiers et aux libertés, garantit aux individus un droit d'accès et de rectification pour les données les concernant. Ce droit peut être exercé auprès des Directions Régionales de l'Insee.

À _____, le _____ 1995
Signature du déclarant :

IMPRIMÉ NUMÉRO **2**

BULLETIN INDIVIDUEL

RECENSEMENT DE LA POPULATION **1996** APPLICATION PILOTE

Cocher la case qui convient ×

I_P 57

(1) NOM, Prénoms *Écrivez le nom en capitales (exemple : ALLARD, épouse MAURIN, Françoise).*

(2) ADRESSE

(3) SEXE Masculin Féminin

(4) ÉTAT MATRIMONIAL LÉGAL

Célibataire Marié(e) Veuf(ve) Divorcé(e)

Une personne vivant en union libre cochera la case correspondant à sa situation juridique présente : si, par exemple, elle est célibataire, elle cochera la 1re case.
Une personne en instance de divorce, ou séparée de son conjoint légitime, cochera la 2e case.

(5) DATE ET LIEU DE NAISSANCE

Né(e) le : *jour* *mois* *année*

à (commune) :

Département :

(pays pour l'étranger, territoire pour les TOM)

Si vous êtes né(e) à l'étranger, en quelle année êtes-vous arrivé(e) en France ? 19 ..

(6) NATIONALITÉ

- **Française**
  - Vous êtes né(e) français(e) .....
  - Vous êtes devenu(e) français(e) par naturalisation, déclaration, à votre majorité ou par manifestation de volonté, etc. .....
    Indiquez votre nationalité à la naissance :
- **Étrangère** .....
  Indiquez votre nationalité :

(7) ÊTES-VOUS INSCRIT(E) pour l'année scolaire 1996-1997 dans un ÉTABLISSEMENT D'ENSEIGNEMENT ?

OUI NON

si oui, l'établissement est-il situé :

- Dans la commune où vous résidez .....
  (ou l'arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une autre commune *(ou un autre arrondissement)*.....
  Indiquez cette autre commune *(précisez l'arrondissement)*

Département : Commune :

(8) OÙ HABITIEZ-VOUS LE 1er JANVIER 1990 ?
(pour toute personne née avant le 1er janvier 1990)

Si, le 1er janvier 1990, vous étiez militaire ou élève interne, indiquez l'adresse de votre résidence personnelle à cette date et non celle de l'établissement (caserne, internat).

- Dans le même logement que maintenant .....
- Dans un autre logement de la même commune .....
  (du même arrondissement s'il s'agit de Paris, Lyon, Marseille)
- Dans une autre commune *(ou un autre arrondissement)* .....
  Indiquez cette autre commune *(précisez l'arrondissement)* :

Commune :

Département :

(pays pour l'étranger, territoire pour les TOM)

Les questions 9 à 24 ne s'adressent qu'aux personnes de 14 ans au plus

(9) QUEL NIVEAU D'ÉTUDES AVEZ-VOUS ATTEINT ?

- École primaire .....
- Collège, classes de 6e à 3e, CAP, BEP .....
- Classes de seconde, première ou terminale .....
- Études supérieures (facultés, IUT, etc.) .....

(10) INDIQUEZ VOTRE DERNIER DIPLÔME OBTENU

- Aucun diplôme .....
- Certificat d'études primaires .....
- BEPC, brevet élémentaire, brevet des collèges .....
- CAP .....
- BEP .....
- Baccalauréat général (Philo., Sciences ex., Math. élém., A, B, C, D, E, L, S, ES, etc.) .....
- Baccalauréat technologique (F, G, H, STI, etc.) ou professionnel, brevet professionnel ou de technicien, autre brevet (BEA, BEC, BEI, etc.), capacité en droit .....
- Diplôme universitaire de 1er cycle, BTS, DUT, diplôme des professions sociales ou de la santé .....
- Diplôme universitaire de 2e ou 3e cycle, diplôme d'ingénieur, d'une grande école, etc. .....

(11) QUELLE EST VOTRE SITUATION ?

VOUS TRAVAILLEZ
Cochez la case et passez au verso (questions 15 à 24) *y compris si vous êtes en congé de* ***maladie*** *ou de* ***maternité,*** *si vous* ***aidez*** *un membre de votre famille dans son travail ou si vous êtes* ***apprenti sous contrat, stagiaire rémunéré*** .....

VOUS NE TRAVAILLEZ PAS (OU PLUS)
Cochez la case et répondez aux questions 12 à 14. .....

(12) ÊTES-VOUS ?

- **Étudiant** *(facultés, IUT, grandes écoles, etc.)* .....
- **Élève** *(collège, lycée)* .....
- **Stagiaire non rémunéré** .....
- **Chômeur** *(inscrit ou non à l'ANPE)* .....
- **Préretraité** .....
- **Retraité :** - ancien **salarié** .....
  - ancien **indépendant** *(agriculteur, artisan, commerçant, etc.)* .
- **Autre** *(femme ou homme au foyer, personne ne percevant qu'une pension de réversion ou d'invalidité, etc.)* .....

(13) CHERCHEZ-VOUS UN EMPLOI ?

- Vous ne cherchez pas d'emploi .....
- Vous cherchez un emploi depuis :
  moins d'un an plus d'un an

(14) AVEZ-VOUS DÉJÀ TRAVAILLÉ ?

OUI → Quelle était votre **profession principale ?**

NON

Signez le bulletin au bas de la page 2

POUR LES PERSONNES DE PASSAGE (voir page 3 de l'imprimé n° 1), **adresse de la résidence habituelle :**

N° Rue (ou lieu-dit) Code postal et commune :

VOUS TRAVAILLEZ

(15) PROFESSION EXERCÉE ACTUELLEMENT

Soyez précis. (Ex. : électricien ***d'entretien de robot****, comptable* ***d'assurances****, technicien* ***chimiste****, etc.)*

Si vous êtes agent de la Fonction Publique de l'État ou des collectivités (y compris HLM, hôpitaux publics), ***précisez votre grade*** *(corps, catégorie...)*

(16) TRAVAILLEZ-VOUS?

- **À temps complet** .....
- **À temps partiel :**

  plus d'un mi-temps — à mi-temps ou moins

Le temps partiel est déterminé par rapport au temps de travail normal dans votre entreprise.

(17) VOTRE ACTIVITÉ PRINCIPALE consiste-t-elle à aider un membre de votre famille dans son travail?

(que vous perceviez ou non un salaire)

(Exploitation agricole ou artisanale, commerce, profession libérale, etc.)

OUI — NON

(18) OÙ TRAVAILLEZ-VOUS?

(a) **Adresse de votre lieu de travail :** *(Ex. : 18, boulevard Pasteur)*

Si travail à domicile, indiquez ***« à domicile »***
Si travail chez un particulier, indiquez ***« particulier »***
Si lieu de travail variable, indiquez ***« variable »***

- Est-ce dans la **commune** où vous résidez ?
  (ou dans l'arrondissement s'il s'agit de Paris, Lyon, Marseille)

  OUI — NON

Si non, indiquez la commune où vous travaillez :
(précisez l'arrondissement)

Commune :

Département :

(pays pour l'étranger)

(b) **Nom (ou raison sociale) de l'établissement qui vous emploie ou que vous dirigez :**

(c) **Adresse de cet établissement,** si elle est différente de celle déclarée à la question 18 a.

(d) **Activité de cet établissement :**
Soyez très précis. *(Ex. : commerce* ***en gros*** *de fruits et légumes, fabrication d'outillage* ***mécanique****, etc.)*

(19) QUEL MODE DE TRANSPORT UTILISEZ-VOUS LE PLUS SOUVENT POUR ALLER TRAVAILLER?

- **Pas de transport** *(travail à domicile)* .....
- **Marche à pied uniquement** .....
- **Un seul mode de transport**

  Deux-roues — Voiture particulière — Transports en commun

- **Plusieurs modes de transport** .....

(20) ÊTES-VOUS?

- **Indépendant ou à votre compte,**
  y compris aide familial non salarié .....
- **Chef d'entreprise salarié,**
  PDG, gérant minoritaire de SARL, co-gérant .....
- **Salarié,** y compris aide familial salarié .....

(21) SI VOUS ÊTES À VOTRE COMPTE OU CHEF D'ENTREPRISE Combien de salariés employez-vous?

Aucun — 1 ou 2 — 3 à 9 — 10 ou plus

Ne comptez ni les apprentis ni les gens de maison. Dans l'agriculture, comptez seulement les salariés permanents.

Les questions 22 à 24 ne s'adressent qu'aux salariés et aux stagiaires rémunérés

(22) INDIQUEZ VOTRE TYPE DE CONTRAT OU D'EMPLOI

- Apprenti sous contrat .....
- Placé par une agence d'intérim .....
- Contrat Emploi Solidarité (CES) .....
- Stagiaire rémunéré (SIFE, etc.) .....
- Contrat de travail à durée déterminée
  (y compris contrat court, saisonnier, etc.) .....
- Emploi à durée indéterminée .....
- Titulaire de la Fonction Publique .....

(23) INDIQUEZ LA CATÉGORIE PROFESSIONNELLE DE VOTRE EMPLOI

- Manœuvre, ouvrier spécialisé (OS1, OS2, OS3, etc.) .....
- Ouvrier qualifié ou très qualifié (P1 à P3, TA, OQ, OHQ, etc.) .....
- Agent de service, aide soignant(e), employé de maison .....
- Employé de commerce, employé de bureau, personnel administratif de catégorie C ou D de la Fonction Publique .....
- Agent de maîtrise dirigeant des ouvriers, maîtrise administrative, commerciale, informatique .....
- Agent de maîtrise dirigeant des techniciens ou d'autres agents de maîtrise .....
- Technicien, dessinateur, VRP .....
- Instituteur, infirmier(e), travailleur social, technicien médical, personnel administratif de catégorie B de la Fonction Publique .
- Ingénieur, cadre d'entreprise *(les techniciens et agents de maîtrise ne devront pas se classer ici, même s'ils cotisent à une caisse de retraite des cadres)* .....
- Personnel de catégorie A de la Fonction Publique et assimilés ..

(24) INDIQUEZ LA FONCTION PRINCIPALE DE VOTRE EMPLOI

- Production, fabrication, chantier, exploitation .....
- Installation, réglage, réparation, maintenance .....
- Gardiennage, nettoyage, entretien ménager .....
- Manutention, magasinage, transports, logistique .....
- Secrétariat, guichet, saisie, standard, accueil .....
- Gestion, comptabilité, fonction administrative, organisation ....
- Directeur général ou adjoint direct, état-major .....
- Commerce, vente, technico-commercial .....
- Recherche, études, méthodes, informatique .....
- Enseignement, formation, santé, travail social, information, publicité, arts, spectacles, sports .....

Nous vous remercions de votre participation

Enquête statistique non obligatoire.
Questionnaire confidentiel destiné à l'INSEE.
La loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés garantit aux individus un droit d'accès et de rectification pour les informations les concernant. Ce droit s'arces peut être exercé pendant le délai où ces informations sont gardées sous forme nominative auprès des directions régionales de l'INSEE.

À _____, le _____ 1996

Signature du déclarant :

PAGE 2

L'INCIDENCE DU CARACTÈRE OBLIGATOIRE DES ENQUÊTES

Catherine Berthier et Françoise Dupont

L'opération méthodologique

Origine de l'opération

Les enquêtes réalisées par l'Insee auprès des ménages ont systématiquement bénéficié jusqu'en 1993 du statut obligatoire (loi n°51-711 du 7 juin 1951 modifiée sur l'obligation la coordination et le secret en matière statistique) quelque soit le thème traité, ce qui constitue une spécificité française dans le paysage statistique international.

En 1993, l'enquête sur les transports, puis en 1994, l'enquête européenne "panel européen" se sont vues refuser successivement ce statut par la Commission Nationale Informatique et Liberté (Cnil). A la suite de ces difficultés, l'Insee a décidé de monter une opération test pour connaître l'impact du statut obligatoire sur la qualité des enquêtes qu'elle réalise.

Si l'idée d'un plan d'expérience paraissait a priori séduisante, très vite la réalisation s'est révélée impossible dans son organisation concrète, sans s'éloigner de conditions de collecte réalistes et donc sans remettre en cause l'extrapolabilité de l'opération. Le paragraphe suivant qui décrit grossièrement les modalités de la collecte auprès des ménages permettra de s'en rendre compte rapidement.

Principes généraux d'organisation des enquêtes ménages à l'Insee

La quasi-totalité des enquêtes que réalise l'Insee auprès des ménages sont effectuées en face à face par un réseau fixe d'enquêteurs répartis sur toute la France et gérés par 18 Directions Régionales. Seules deux enquêtes sont réalisées par téléphone : l'enquête emploi trimestrielle qui constitue la continuation auprès des mêmes ménages de l'enquête emploi annuelle réalisée en face à face, et l'enquête de conjoncture atypique par le choix des questions et la faible durée d'interrogation.

Pour les enquêtes en face à face, la saisie portable se développe mais la majorité des enquêtes est encore réalisée sur support papier.

Pour chaque enquête, des logements sont sélectionnés dans une réserve selon la technique de l'échantillon maître. L'équivalence entre résidences principales et ménages permet d'obtenir un échantillon de ménages. Les ménages sélectionnés font l'objet d'une ou plusieurs visites (trois maximum) d'une heure en moyenne. Les questions sont le plus souvent relatives au ménage dans son ensemble, mais détaillent parfois les membres du ménages et/ou se concentrent sur un individu tiré au hasard selon la technique du tirage Kish. Selon les enquêtes, les questionnaires sont complétés par des carnets où le ménage inscrit entre deux visites de l'enquêteur ses dépenses (enquêtes budget), ses déplacements (enquête transport) etc...

Le résultat du tirage de l'échantillon effectué de manière centralisée est transmis aux Directions Régionales (DR) qui sélectionnent les enquêteurs pour l'opération, et répartissent les logements de l'échantillon entre les enquêteurs. Elles préviennent les autorités locales (mairie, commissariat, préfecture, gendarmerie) selon des modalités différentes selon les DR.

Les enquêteurs effectuent, avant la date de début de collecte, un repérage sur le terrain des logements qui leur ont été attribués sur la base d'adresses extraites du recensement, ou, s'il s'agit d'un logement construit après le recensement, sur la base du suivi des permis de construire.

Ils déterminent les logement vacants ou résidences secondaires, lorsque l'information est disponible à ce moment de l'enquête, pour tous les logements considérés comme à enquêter (incluant les logements au statut incertain). Un courrier annonçant la réalisation d'une enquête, et la venue prochaine de l'enquêteur est alors envoyé sur la base des renseignements collectés par l'enquêteur.

Une fois la période de collecte préconisée par le concepteur commencée, l'enquêteur prend rendez-vous avec le ménage. Bien que la visite au domicile soit préconisée, le nombre croissant d'interphone et la faible disponibilité des ménages incitent les enquêteurs à recourir au téléphone pour fixer un entretien (pour un tiers dans l'enquête).

Plusieurs passages au domicile ou essais téléphoniques sont souvent nécessaires pour obtenir un contact avec une personne du ménage. Seules les personnes majeures du ménage sont habilitées à répondre pour le ménage. Ainsi les portes fermées, la présence des seuls enfants ou amis conduira l'enquêteur à réitérer ses essais. Si le ménage exprime explicitement un refus, une lettre de relance pour « refus » lui sera envoyée par la DR sur demande de l'enquêteur. Si le ménage s'avère impossible à joindre après quelques essais (y compris si le ménage ne répond pas alors qu'il est manifestement présent au domicile), une lettre de relance

pour « impossible à joindre » lui sera envoyé par la DR sur demande de l'enquêteur. Dans ces deux lettres assez proches dans leur contenu, il est demandé au ménage de fixer une date de rendez-vous avant une date limite. Dans certains cas (refus obstiné ou difficultés du ménage), les circonstances décrites par l'enquêteur amène la DR à renoncer à une relance.

A l'issue de la relance par lettre, diverses situations sont possibles : pas de réponse à la lettre, réponse proposant un rendez-vous une fois le délai dépassé, réponse négative à la lettre, passage de l'enquêteur avant réponse à la lettre, appel de l'enquêté à la DR et rendez-vous fixé par la DR.

Le résultat de l'enquête sera considéré comme définitif après un ultime essai de l'enquêteur, au cours duquel il peut ou non y avoir contact. Un refus peut être exprimé à ce stade.

Cette description rapide fait apparaître plusieurs difficultés dans la mesure des phénomènes de réponse du ménage

- Le processus d'interaction entre l'enquêteur et le ménage est itératif, le nombre d'itérations est variable. Les intervenants du ménage aux différentes étapes sont parfois différents. La mesure des phénomènes ne peut pas toujours se faire sur toutes les itérations. Elle se focalise nécessairement sur des étapes importantes.

- Une partie de l'information sur le résultat de collecte se trouve au niveau de la DR (lettre retournée avec la mention n'habite pas à l'adresse indiquée, refus téléphonique de l'enquêté, résultat de la relance). Il faut donc s'assurer lorsqu'on demande le résultat à l'enquêteur que cette information lui est systématiquement retournée.

- Enfin les frontières entre certaines situations sont délicates à définir sans ambiguïté : le ménage qui reste derrière sa porte doit-il être comptabilisé comme impossible à joindre ou comme un refus (déguisé)?. La personne de 80 ans manifestement en bonne santé qui parle de son âge pour échapper à l'enquête est-elle inapte à l'enquête ou bien doit-on considérer qu'elle refuse l'enquête ? Des définitions aussi précises que possible ont été adoptées pour définir les différentes catégories de ménage et comptabiliser les refus. Toutefois des difficultés subsistent, comme nous l'a montré par la suite l'exploitation détaillée, pour délimiter des ménages dits hors champ et des ménages impossible à joindre.

Que mesurer dans l'opération ?

L'hypothèse a priori, appuyée par des dires d'enquêteurs et les taux de réponses observés pour l'enquête panel européen, était que l'obligation permettait d'obtenir

des taux de réponse plus élevés, et donc une meilleure qualité. L'objectif principal de l'opération était donc de mesurer l'effet de l'obligation sur les taux de réponse.

Etant donné l'émergence récente du label d'intérêt général, une question annexe était le degré de substituabilité entre l'obligation et le label d'intérêt général.

La dégradation de la qualité de l'enquête due à des réponses approximatives ou fausses de la part des enquêtés contraints de répondre est difficile, voir impossible à cerner. L'hypothèse que nous avons faite, fondée sur les dires d'enquêteurs, est que cet effet est du second ordre. Les enquêteurs préfèrent mettre fin à l'entretien lorsqu'ils constatent un comportement de réponse sciemment erronée. Cet aspect n'a donc pas été privilégié dans l'opération. Des réponses ont toutefois été collectées et saisies. Une analyse de la dispersion des réponses est donc possible et peut permettre de vérifier la présence d'un comportement de réponse anormalement différent. Toutefois il s'agit là d'un travail important au regard du résultat probable d'absence d'effet significatif.

Le comportement des ménages face à l'obligation étant mal connu, on a souhaité recueillir à l'occasion de cette opération des renseignements complémentaires permettant de mieux comprendre l'impact de l'obligation au niveau de chaque ménage. Ce point sera détaillé par la suite.

Le choix du dispositif de base

La principale difficulté résidait dans l'objet même de l'observation. L'obligation de réponse intervient pour un même ménage à différent moments de la collecte et une grande variété de situations sont possibles. On dispose de peu d'information quantitative décrivant l'organisation de la collecte qui relève pour une part de l'initiative des directions régionales et dont les paramètres ne sont pas aisément contrôlables. Enfin, tous ceux qui ont fait des enquêtes le savent bien, les facteurs influençant le taux de réponse sont nombreux.

L'analyse de l'existant en collaboration étroite avec les directions régionales (personnel des DR chargé des enquêtes et enquêteurs) impliquées dans l'opération a donc constitué une étape importante dans la préparation de l'opération. De même le bilan de collecte effectué avec une partie des enquêteurs nous a permis de valider le dispositif et de tirer quelques enseignements importants. Il s'en est dégagé l'impression que les enquêteurs étaient moins négatifs a posteriori sur le statut facultatif.

Si l'idée d'un plan d'expérience paraissait s'imposer au début de la conception de l'opération, l'analyse détaillée conduisait progres-sivement vers plus de pragmatisme en raison des contraintes liées aux nombreux facteurs annexes.

L'interrogation des ménages sur un comportement potentiel de réponse a été éliminé d'emblée comme non pertinent. L'utilisation de plusieurs sujets d'enquête aurait nécessité des formateurs pour chacune de ces enquêtes, et risquait d'entraîner une surcharge pour chaque enquêteur (problème de documents et d'argumentaires différents). Il fallait limiter à une taille raisonnable la gestion de courriers différents pour chaque enquête en DR, ainsi que le nombre de questionnaires. Ces problèmes d'organisation et la contrainte de la taille de l'échantillon nous ont conduit à renoncer à employer plusieurs enquêtes.

Les enquêtes candidates pour l'opération étaient alors des enquêtes réalisées en face à face avec un questionnaire papier (CAPI pose des problèmes spécifiques dans la mention de l'obligation), et récentes en raison des problèmes de logistique. De plus, il semblait préférable de retenir un enquête à un seul passage, dont le sujet soit favorable en matière de réponse. Les bilans de collecte ont relativisé l'importance de ces deux derniers points. En effet, durant le bref échange qui permet à l'enquêteur d'obtenir la participation de l'enquêté, le sujet n'est présenté que très schématiquement, surtout si l'enquête est multithème, et le nombre de passages nécessaires n'est le plus souvent pas évoqué.

Le choix s'est porté sur une enquête en cours aux dates de l'opération, l'enquête permanente sur les conditions de vie (P.C.V.). Ce choix nous a permis de profiter d'une collecte déjà menée pour accroître la taille de l'échantillon utilisable pour l'opération et de simplifier les problèmes d'organisation de la formation. Les facteurs de charge d'enquête, de sujets traités, de durée annoncée de l'entretien étaient alors figés.

Le choix de l'enquête réelle P.C.V. comme point de référence de la situation avec obligation de réponse nous conduisait alors naturellement à deux échantillons indépendants pour les deux situations obligatoire et facultative :

- l'échantillon résultant sur les trois DR de l'échantillon tiré nationalement pour l'enquête P.C.V.

- un échantillon tiré spécifiquement pour l'opération, pour servir de point de référence de la situation facultative appelé PCVM en interne.

Les moyens alloués à l'opération nécessitait de se restreindre à un petit nombre de DR participantes. Un appel d'offre a été lancé. Le choix s'est donc effectué en fonction des disponibilités des DR. Compte tenu des problèmes particuliers que rencontrent les enquêteurs en Ile de France, nous souhaitions vivement la participation de la DR de St-Quentin, qui a pu effectivement participer à l'opération. Les DR de Champagnes Ardennes, Languedoc Roussillon et Paris sélectionnées présentent des différences en matière d'organisation de la collecte, de réseau d'enquêteur, de terrain d'enquêtes. La répartition sur trois DR devait permettre de

s'affranchir de l'effet DR (terrain et organisation) pour garantir une extrapolabilité correcte.

Le tirage de l'échantillon dans l'échantillon maître s'est imposé pour plusieurs raisons. La première tient à la disponibilité des enquêteurs sur les zones de l'échantillon. Les enquêteurs travaillaient dans leur conditions habituelles en matière de connaissance des terrains qu'ils leur étaient affectés. Enfin il permettait dans les zones rurales d'éloigner les zones " obligatoires " et " facultatives " afin de limiter les possibilités de communication entre personnes ayant connaissance de statuts d'enquête différents. Si la transparence sur l'opération a été la règle en interne, l'opération était présentée comme une enquête ordinaire aux autorités locales (mairie, commissariat, gendarmerie, etc...)[1]. Les risques de dérapages sur le statut de l'enquête liés à l'intervention des autorités locales ou à une communication entre enquêtés, qui nous posaient problème a priori, ne se sont d'ailleurs pas produits.

Outre l'effet terrain, le risque le plus important étant probablement l'effet enquêteur, nous souhaitions nous en affranchir autant que les contraintes pratiques nous le permettaient. Le principe était donc que chaque enquêteur réalise des enquêtes obligatoires et des enquêtes facultatives. En fonction du résultat du tirage, les allocations ont été réalisées de manière aussi équilibrée que possible entre les deux versions facultative et obligatoire.

L'équivalence des périodes de janvier et de février (à l'exclusion des vacances scolaires) permettait d'envisager deux collectes disjointes dans le temps pour les deux versions facultative et obligatoire. Cette option minimisait les risques d'erreurs entre statuts lors des appels des enquêtés à la DR. Elle permettait en outre d'éviter les confusions entre documents au niveau des enquêteurs et d'éviter la gymnastique entre deux argumentaires un même jour pour des enquêtés différents. Cette précaution s'est révélée très pertinente a posteriori. On a donc défini des périodes de collectes semblables non seulement dans leur durée, mais aussi dans le nombre de week-end inclus et dans le jour de la semaine constituant le premier jour. Des dates de collecte précises (permettant la formation, situées hors période scolaire, avec une répartition équilibrée des jours d'indisponibilité de l'enquêteur) ont été définies avec chaque DR. Les fins de collectes ont donc été gérées plus strictement que dans une collecte habituelle.

Une fois fixée l'organisation générale, celle de deux collectes, réalisées par les mêmes enquêteurs, sur deux échantillons de *ménages* [2] de taille équivalente, dans

---

1 En cas de question, dans les quelques petites communes à deux statuts, la règle était un statut obligatoire par défaut. CNIS et CNIL avaient été informés de l'opération.

2 le concept retenu était un taux de réponse de ménages.

trois DR, sur deux périodes de temps disjointes et équivalentes, avec le même questionnaire support, il restait à définir plus précisément les modalités de la collecte, en particulier en matière d'obligation.

En matière de collecte quelques spécificités des régions ont été conservées. Les textes de lois ont été joints à la lettre-avis à Paris. Quelques appels téléphoniques ont été effectués par la DR de Languedoc Roussillon, ainsi que quelques rappels en recommandé.

Pour la mention de l'obligation par l'enquêteur, les consignes usuelles ont été reprises : pas de mention explicite systématique de l'obligation, utilisation en dernier recours de cet argument.

Le problème des courriers envoyés au ménage était plus délicat. Les courriers adoptés devaient être les mêmes dans les trois régions, en même temps qu'ils devaient être proches des rédactions les plus répandues.

La lettre-avis annonçant le statut de l'enquête a fait l'objet de soins particuliers. Cependant, l'opinion des enquêteurs qui indiquent que la lettre-avis est peu lue s'est vue confirmée par l'enquête. L'examen des lettres avis des enquêtes passées des 18 DR a montré une diversité de solutions adoptées tant en matière d'obligation qu'en matière d'argumentaire. L'obligation est en effet, dans la première lettre avis envoyée au ménage, tantôt mentionnée dans le corps de la lettre, tantôt mentionnée en note de bas de page, tantôt non mentionnée; Les lettres de rappel sont plus homogènes et font figurer le plus souvent l'obligation, pour la plupart dans les mêmes termes.

Le caractère facultatif n'était jusqu'à présent mentionné dans aucun courrier. L'évolution de la position de la Cnil en faveur d'une information destinée à l'enquêté, dont le texte lui serait soumis, nous a conduit à adopter pour chacun des deux statuts une mention explicite dans tous les types de courriers. Les bilans de collecte ont toutefois relevé que les relances explicitant le caractère facultatif ne sont pas opérationnelles et surprennent le ménage.

En revanche toutes les enquêtes bénéficiant du label d'intérêt général en font mention dans leurs courriers. Cette mention a donc figuré dans les lettres, pour les deux collectes obligatoire et facultative. On a ensuite cherché à savoir si, du côté facultatif, le label d'intérêt général jouait un rôle comparable à l'obligation dans la décision de participation.Une partie des questionnaires déposés auprès des ménages a été construite de façon symétrique pour les deux versions, de façon à savoir si le label jouait le rôle de substitut à l'obligation. Bien que les libellés des questions aient été choisis précisément dans ce but, l'impact du label est difficile à mesurer. Il apparaît que 62% des ménages ont participé à l'enquête facultative « parce qu'elle était reconnue d'intérêt général ». Toutefois ce résultat doit être pris avec prudence

car il ressort des bilans de collecte que le label, argument essentiellement écrit, se confond par oral avec un discours sur l'intérêt de l'enquête : La frontière est trop ténue entre une enquête présentée d'intérêt général en raison de son contenu et une enquête « labelisée » d'intérêt général.

Le dispositif complémentaire

Afin d'obtenir des éléments d'explications sur l'impact de l'obligation au niveau du ménage, deux brefs questionnaires complémentaires ont été réalisés :

- un questionnaire rempli par l'enquêteur recueillant des informations sur l'utilisation de l'obligation, les circonstances du refus ou de l'acceptation et permettant d'identifier les différents interlocuteurs de l'enquêteur

- un questionnaire rempli après le départ de l'enquêteur puis renvoyé à la DR par le ménage ayant répondu à l'une des enquêtes obligatoire ou facultative,.

Le premier questionnaire était destiné au départ à comptabiliser les relances au niveau global, ce qui n'était pas réalisé jusqu'à présent, et à repérer les ménages relancés.

Le deuxième questionnaire avait pour objectif prioritaire d'éclairer le comportement des répondants à l'enquête obligatoire : Il s'agissait de déterminer le pourcentage et le profil des ménages participant aux enquêtes en raison de l'obligation. D'autres renseignements ont été collectés à cette occasion pour ne pas compromettre ce questionnement et pour disposer en outre d'informations sur la perception du déroulement de la collecte.

Pour ce dernier questionnaire, plusieurs pistes différentes ont été explorées.

Compte tenu de l'objet de l'interrogation (obligation et conditions de collecte), nous ne souhaitions pas d'interaction entre l'enquêteur et l'enquêté dans le questionnement.

Le questionnement téléphonique par des agents de la DR posait plusieurs problèmes. Il fallait joindre au sein du ménage une ou deux personnes précises, celles qui avaient participé à l'enquête. Il fallait respecter un délai de réflexion entre l'enquête proprement dite et ce nouveau contact. Enfin, l'interrogation téléphonique demandait une disponibilité du personnel en DR à des heures adaptées.

Nous avons donc abandonné la perspective d'un questionnement téléphonique. Or le mode postal ne donne traditionnellement pas un taux de réponse très élevé. Nous avons finalement utilisé un mode de collecte mixte original : dépôt par l'enquêteur

chargé de motiver l'enquêté sur le questionnaire sans l'influencer sur la réponse à donner, retour postal et rappel téléphonique par les DR, au bout de dix jours. Un test réalisé à Montpellier avait permis de mettre au point les modalités pratiques de ce questionnement. Les efforts conjugués des enquêteurs et des gestionnaires ont permis d'atteindre dans les trois régions un taux de réponse record de 86% permettant l'exploitation statistique du questionnaire.

Ce mode de questionnement particulier a entraîné des difficultés de mise en oeuvre. Six versions du questionnaire ont été élaborées pour coller au différentes situations (rappel ou non, refus une première fois ou non, obligatoire ou facultatif). Le pointage des retours des questionnaires renvoyés par les ménages, les rappels, et la fusion avec le dossier retourné par l'enquêteur ont dû être coordonnés en DR au prix d'une charge de travail importante.

En outre, nous avons exploré la piste d'une interrogation des non répondants. Un micro test a été effectué. Comme on pouvait s'y attendre, le mode postal induit des réponses extrêmes. Le mode téléphonique se heurte à de nombreuses difficultés, dont celle de joindre la personne ayant refusé. Les déclarations de mauvaise foi sont fréquentes " je ne suis pas madame X, je n'ai pas vu l'enquêteur etc ". Cette interrogation a donc été abandonnée.

En guise de conclusion sur l'organisation

Il me paraît important de souligner, pour conclure sur l'organisation, que la principale difficulté de ce type d'opération méthodologique accroché à une collecte existante est l'impossibilité par nature à tester le dispositif dans sa totalité.

Les premiers résultats de l'enquête Obligation

Un impact évident de l'obligation

Le résultat de l'opération est sans ambiguïté puisque dans les trois régions le taux de refus à la fin de l'enquête double lorsque l'enquête devient facultative : passant de 12% à 23% pour Paris et la petite couronne, de 8 à 18 % en Languedoc Roussillon, et de 7 à 18% en Champagne, soit 9% à 19% pour l'ensemble des régions.

On peut penser pallier de façon simple, quoique coûteuse, la réduction de l'échantillon de réponses qu'entraîne le passage au statut facultatif. Il pourrait suffire d'enquêter un plus grand nombre de ménages. Mais ce changement de statut induit également, et c'est le plus important, une déformation de structure de la

population qui répond. Or cette déformation ne pourra être au mieux que partiellement compensée par les redressements effectués par les statisticiens. Ces redressements demandent en outre de connaître les spécificités des ménages non-répondants.

Une première approche, pour comprendre l'impact de l'obligation, consiste à comparer, dans chacune des deux enquêtes, les ménages donnant leur accord et ceux opposant un refus. Mais, sur ces derniers, aucun renseignement n'a pu être collecté. Il est seulement possible d'utiliser les données rattachées au logement datant du recensement de 1990. Selon cette source, les comportements de refus apparaissent nettement différenciés selon l'âge (il s'agit de l'âge de la personne de référence du ménage). Les jeunes (moins de 35 ans) réagissent le plus fortement au changement de statut de l'enquête, le taux de refus passant de 6% pour l'enquête obligatoire à 20% pour celle facultative. Il n'empêche que les refus deviennent plus fréquents à mesure que l'âge s'accroît. Les plus âgés (plus de 65 ans) opposent, dans les deux cas, les taux de refus les plus élevés (18% en obligatoire et 31% en facultatif). Ces résultats sont fragiles, puisqu'ils tablent en cas de changement d'occupant du logement, sur la permanence de l'âge.

Cet effet de l'âge doit être nuancé, parce qu'il recoupe celui de la région. Ainsi, l'impact du passage d'obligatoire à facultatif sur les refus des jeunes (moins de 35 ans) est en fait limité à Paris et la petite couronne. Il est amplifié par le fait que la région parisienne est la plus jeune. Mais il est à examiner comme un comportement des jeunes parisiens, en raison des spécificités de ce groupe. En Champagne, le changement de statut de l'enquête touche les ménages d'âge médian, plus représentés parmi les refus de l'enquête facultative.

Ces effets sont difficiles à corriger, et même à soupçonner en examinant la structure des répondants. Il se trouve que pour ces deux caractéristiques, l'âge et la région, une fois rendus équivalents les échantillons des deux enquêtes, les répondants présentent des structures tout-à-fait semblables, alors que des différences sont visibles sur les structures des refus. Ce paradoxe vient à point nommé pour rappeler que certaines des déformations induites par le changement de statut de l'enquête seront invisibles sur les seuls répondants.

Pour chacune des deux versions de l'enquête, un ensemble de documents méthodologiques ont été collectés, retraçant les points de vue de l'enquêteur et du ménage répondant. L'enquêteur a dû rendre compte des conditions dans lesquelles se sont déroulés les contacts avec le ménage; l'enquêté, s'il a accepté de participer, a été sollicité après coup sur ses motifs de participation.

A propos des refus, on dispose donc des renseignements fournis par l'enquêteur. Pour les entretiens réalisés, chacun de leur côté, l'enquêteur et l'enquêté se sont

exprimés sur les conditions dans lesquelles l'accord a été obtenu, et précisément sur le rôle qu'a joué le statut de l'enquête.

D'une version à l'autre de l'enquête, des motifs de participation semblables au premier contact ...

Les motifs de refus restitués par les enquêteurs se répartissent de la même manière dans les deux enquêtes: l'hostilité au sondage et le manque de temps sont cités par plus de 20% des ménages, les raisons personnelles et l'âge, l'intrusion dans la vie privée sont cités par environ 10% des ménages chacun. L'hostilité à l'état, l'indiscrétion des sujets, le manque d'intérêt pour les sujets sont peu cités (moins de 5% chacun). En revanche, le refus est moins souvent exprimé avec courtoisie et plus souvent avec méfiance ou brutalité lorsque l'enquête est obligatoire. Le jugement des enquêteurs est sévère puisque dans 15% des cas seulement ils jugent ces motifs justifiés et s'ils n'ont pas d'opinion pour 20% des refus, ils jugent qu'il s'agit d'un manque d'effort de l'enquêté dans 60% des cas.

Quel que soit le statut de l'enquête, l'enquêteur joue le même rôle pour la faire accepter d'emblée, sans relance : Malgré la lettre avis, environ 20% des ménages répondants n'acceptent qu'après argumentation de l'enquêteur, dont 5% difficilement en raison de fortes réticences.

Dans le cadre facultatif, les refus impliquent une charge plus lourde de relances. Les relances par lettres aboutissent plus souvent à la réalisation de l'enquête, mais sont loin de compenser les échecs au premier contact. En fait, dans l'enquête facultative, cette forme de relance a été moins souvent tentée : seulement un quart des ménages qui en bout de course ont refusé de répondre ont fait l'objet d'une relance par lettre, contre un peu plus de la moitié dans l'enquête obligatoire.

Tableau 1 : Les relances par lettre pour refus

	Enquête obligatoire	Enquête facultative
Enquête réalisée	1318	1149
dont relance pour refus	57	157
Enquête refusée	169	362
dont relance	93	84

... motifs sur lesquels les enquêtés se sont exprimés après coup

Il était important, dans une telle opération, d'avoir une idée des raisons poussant les personnes à répondre à l'enquête. Le questionnaire d'une page déposé par

l'enquêteur a été renvoyé par 86% des ménages. Les enquêtés de la version facultative se sont d'ailleurs encore plus fortement mobilisés que ceux de l'enquête obligatoire.

Dans les deux enquêtes, les personnes qui ne font pas l'effort de répondre au questionnaire enquêté ont des caractéristiques proches. On perd en priorité les personnes ayant des difficultés de langue. Les non-diplômés répondent moins fréquemment, et ce phénomène est accentué dans l'enquête facultative.

Alors que les motifs de refus ont été exprimés de la même manière au premier contact avec l'enquêteur quelque soit le statut, après coup, les déclarations des enquêtés sont bien distinctes d'une version de l'enquête à l'autre.

On a pu ainsi constater un réaction plus négative vis-à-vis de l'obligation quand on est dans le cadre facultatif : 19% des « facultatifs » auraient été choqués par l'obligation, mais elle n'a choqué que 11% des « obligatoires ». De même, si 37% des ménages « obligatoires » affirment avoir participé en raison de l'obligation, 18% seulement des enquêtés « facultatifs » l'auraient trouvé normale.

Le rôle de l'obligation sur la sélection des répondants

En comparant directement les répondants de chacune des deux enquêtes, on constate des modifications touchant à la composition des ménages. A l'enquête facultative, la proportion de ménages incluant un couple est plus forte; les personnes seules, les ménages sans enfants, les ménages à un enfant sont sous-représentés. Cette déformation peut être rapprochée de celle qu'entraîne la non-réponse (et pas seulement le refus) dans les enquêtes en général. Par exemple, le contact avec un membre d'une famille est plus facile, ne serait-ce que parce que le logement est plus « largement » occupé. Le passage au statut facultatif pourrait faire perdre en priorité les ménages les plus difficiles à atteindre dans toutes les enquêtes.

Un modèle logit opposant les répondants des deux enquêtes met en évidence la différence entre Paris et les deux autres régions de l'échantillon.. Le fait que le ménage soit de petite taille a une portée propre, pas seulement liée à la présence de Paris. On relève aussi un impact de l'âge du chef de ménage. « Toutes choses égales par ailleurs, que le chef de ménage soit âgé de plus de cinquante ans le place plutôt parmi les répondants « facultatifs ». Par contre, ni sa catégorie sociale, ni son statut matrimonial ne paraissent reliés à cette opposition.

Sur la seule enquête obligatoire, plusieurs questions posées à la fois à l'enquêteur et à l'enquêté visent à mesurer le rôle du statut obligatoire sur la décision de participation. Les répondants « obligatoires » peuvent être répartis en deux groupes, suivant que le statut de l'enquête a joué ou non. On est tenté d'examiner ce qui

sépare ces deux groupes, parce qu'on présume que les répondants influencés par l'obligation ressemblent à ceux que l'on perd dans l'enquête facultative.

Tout d'abord, le statut de l'enquête n'est pas toujours perçu par l'enquêté : 17% des ménages « obligatoires » ne l'ont pas remarqué du tout, et 20% l'ont remarqué grâce à l'intervention de l'enquêteur. Pourtant le choix avait été fait de mentionner clairement le statut de l'enquête dans tous les courriers.

Tableau 2 : Modèle logit opposant les répondants de chacune des deux enquêtes

	Coefficient	*Student*
Nombre de personnes du ménage		
- 1	0,32	*1,8*
- 2 ou 3	0,28	*2,2*
- 4	ref	
- 5 ou plus	-0,17	*-1,0*
Taille de l'agglomération		
-petite	-0,32	*-3,4*
-moyenne ou grande	-0,30	*-2,6*
-Paris	ref	
Age du chef de ménage		
- de 20 à 29 ans	0,01	*0*
- de 30 à 39 ans	-0,05	*-0,4*
- de 40 à 49 ans	ref	
- de 50 à 59 ans	-0,29	*-2,2*
- de 60 à 69 ans	-0,23	*-1,6*
- plus de 70 ans	-0,07	*-0,5*
Est-il marié ?		
- oui	ref	
- non	0,01	*0*

Lecture : le coefficient -0,29 associé à la catégorie d'âge de 50 à 59 ans signifie qu'un ménage qui ne diffère de la situation de référence que par l'âge a une probabilité plus faible d'appartenir à l'ensemble des répondants de la version obligatoire.

La statistique de Student permet de vérifier la significativité du coefficient : lorsqu'elle est supérieure ou égale à 2 en valeur absolue, le coefficient est significatif au seuil de 5%.

Pour distinguer parmi les répondants ceux pour lesquels l'obligation a été déterminante, deux questions proches ont été posées à la fois à l'enquêté et à l'enquêteur. La confrontation des réponses laisse apparaître des désaccords. De plus, l'influence de l'obligation apparaît sous deux jours opposés : déclarée après coup par l'enquêté, elle paraît beaucoup plus forte que celle restituée par l'enquêteur

(44% des enquêtés déclarent avoir répondu par obligation, mais pour l'enquêteur seulement 21% des enquêtés ont été influencés par l'obligation).

Tableau 3 : L'influence du caractère obligatoire

498 oui, l'enquêté déclare que l'obligation a joué		
99 accords oui pour enquêté oui pour enquêteur	**235** désaccords oui pour enquêté non pour enquêteur	**164** oui pour enquêté l'enquêteur est indécis
396 non, l'enquêté déclare que l'obligation n'a pas joué		
190 accords non pour enquêté non pour enquêteur	**80** désaccords non pour enquêté oui pour enquêteur	**126** non pour enquêté l'enquêteur est indécis

La question posée à l'enquêteur est : « Selon ce qui a été exprimé spontanément par l'enquêté lors des différents contacts, le caractère obligatoire de l'enquête a-t-il finalement influencé l'acceptation ».

Celle posée à l'enquêté est : « Est-ce parce qu'elle était obligatoire que vous avez participé à l'enquête ? ».

Pour un petit nombre de désaccords, l'enquêté répond qu'il a participé en raison de l'obligation, ... mais également qu'il n'a pas remarqué, avant de commencer à répondre à l'enquête, que celle-ci était obligatoire. On retrouve la tendance chez l'enquêté à adhérer naturellement au statut de l'enquête présentée.

Une fois ces contradictions résolues en partie grâce à d'autres questions, la participation d'un peu plus du tiers des enquêtés « obligatoires » paraît influencée par le caractère obligatoire. Cette proportion est forte au regard du fait qu'on perd environ une personne sur dix en passant au statut facultatif. De plus, c'est à Paris (et petite couronne) qu'on rencontre le plus « d'obligatoires » indifférents à l'obligation, alors que l'enquête facultative s'y heurte au plus fort taux de refus. L'effet de la région est d'ailleurs l'inverse de celui observé en séparant les répondants des deux enquêtes.

Tableau 4 : Répartition des répondants de l'enquête obligatoire

Enquête obligatoire	Paris et petite couronne	Languedoc et Champagne
Répondants par obligation	183	292
Indifférents à l'obligation	533	303
Ensemble	1057	595

L'appartenance des enquêtés « obligatoires » à l'un de ces deux groupes peut être recoupée avec quelques indicateurs sociaux attachés au ménage enquêté, tels que l'âge et la catégorie sociale du chef de ménage; le fait que celui-ci soit marié, le nombre d'enfants et le type d'habitat urbain ou rural. Un modèle logit distingue en tout premier lieu le type d'habitat, qui recoupe l'opposition entre Paris et sa proche banlieue, et les deux autres régions de notre échantillon. Cette fois-ci, l'effet est inverse de celui observé en séparant les répondants selon le statut de l'enquête : Habiter hors Paris fait pencher du côté de l'influence de l'obligation. La catégorie sociale est le second facteur à retenir : « Toutes choses égales par ailleurs », le groupe des employés, et dans une moindre mesure les professions intermédiaires se situent plus du côté de l'influence de l'obligation que les ouvriers. Un nombre élevé d'enfants fait pencher également vers l'obligation.

Tableau 5 : Modèle logit distinguant, parmi les répondants à l'enquête obligatoire, ceux influencés par l'obligation et les « indifférents »

	Coefficient	***Student***
Nombre d'enfants		
- pas d'enfant	0,03	*0,3*
- 1 ou 2	ref	
- 3 ou plus	0,34	*1,6*
Taille de l'agglomération		
- petite	0,95	*6,7*
- moyenne ou grande	0,91	*5,5*
- Paris	ref	
Catégorie sociale du chef de ménage		
- Indépendants (agriculteurs, ou artisans commerçants et chefs d'entreprise)	0,21	*1*
- Cadres	0,01	*0,5*
- Professions intermédiaires	0,15	*0,6*
- Employés	0,40	*2,2*
- Ouvriers	ref	
- Sans activité	-0,23	*-0,5*
Est-il marié ?		
- oui	ref	
- non	-0,14	*-1,1*

Lecture : le coefficient 0,40 associé aux employés signifie qu'un ménage qui ne diffère de la situation de référence que par la catégorie sociale employé a une probabilité plus forte d'être influencé par l'obligation. La statistique de Student permet de vérifier la significativité du coefficient : lorsqu'elle est supérieure ou égale à 2 en valeur absolue, le coefficient est significatif au seuil de 5%.

On peut chercher à rapprocher les « indifférents » de l'enquête obligatoire de l'ensemble des répondants « facultatifs ». Un modèle logit calqué sur le précédent, opposant les enquêtés sensibles à l'obligation aux indifférents auxquels on joint cette fois les répondants facultatifs, livre des résultats tout-à-fait similaires, marqués par la prépondérance du clivage entre Paris et le reste de l'échantillon, et l'opposition entre les employés et les autres professions.

La différence la plus sensible se situe finalement entre les « obligatoires » indifférents au statut, et les autres répondants des deux versions. Elle est liée à l'influence de Paris, prépondérant dans ce groupe des indifférents; on retrouve donc dans ce groupe des personnes plus jeunes, et des ménages de taille réduite.

... à suivre ...

Le plus marquant des résultats est l'évolution des taux de refus d'une version à l'autre. Mais examiner ces refus au travers des questions fermées sur les motifs de participation, ou les voir « en négatif », en étudiant la sélection des répondants ne permet de pas de dégager de cohérence à un niveau qui ne soit plus individuel. Aussi, cette étude sera poursuivie, tournée sur l'exploitation des questions ouvertes, recueillies à plusieurs étapes des questionnaires, du côté enquêteur et enquêté.

L'ensemble des travaux présentés ont été menés en collaboration avec Bernard Neros.

D'autre part, l'exploitation de l'enquête a bénéficié des études réalisées lors de deux stages effectués par Laetitia Bonnani, étudiante de l'université Paris-Dauphine, et par Marina Gatignol, étudiante de l'Ensai.

Session 2

Les séries temporelles

LA DESAISONNALISATION : DES ORIGINES JUSQU'AUX NOUVEAUX LOGICIELS X12-ARIMA ET TRAMO-SEATS

Ketty Attal

Introduction

Le présent document avait initialement pour projet de présenter deux nouveaux logiciels de désaisonnalisation : X12-ARIMA et TRAMO-SEATS. Cependant, pour une meilleure compréhension de ces méthodes, il paraissait utile d'expliquer les évolutions en matière de désaisonnalisation qui ont conduit à celles-ci. Une première partie sera donc consacrée à une présentation générale des méthodes de désaisonnalisation, du point de vue historique et du point de vue de leur philosophie. La deuxième partie fera une description des deux nouveaux logiciels, qui reposent sur des principes très différents, mais qui ont en commun d'une part l'utilisation des modèles ARIMA, d'autre part le souci d'améliorer la "qualité" de la série qui doit être décomposée, par un traitement préalable assez poussé.

Un peu d'histoire

L'histoire des séries temporelles prend source dans l'Antiquité, avec l'observation des mouvements des astres. D'abord purement descriptive, l'étude des séries temporelles s'est ensuite enrichie d'une volonté d'analyse. Celle-ci consiste à isoler les différents mouvements qui entrent en jeu dans le mouvement général de la série. La désaisonnalisation, à savoir l'élimination des variations systématiques (saisonnières) dans la série est un aspect de l'analyse des séries temporelles dont le développement a donc été étroitement lié aux progrès de celle-ci.

Les origines

Dès le 17ème siècle, des astronomes ont fait remarquer qu'une série chronologique pouvait être considérée comme étant formée de composantes inobservables. La première formalisation de la décomposition d'une série remonte aux travaux de Fourier (1807) sur la décomposition d'une série périodique en une somme de fonctions trigonométriques. En 1847, Buys Ballot, météorologue, a recherché les

variations périodiques dans les séries et les estimait grâce à une analyse fondée sur un tableau dans lequel les données étaient arrangées en 12 (resp. 4) colonnes, pour chaque mois (resp. trimestre). Il modélisait la tendance par un polynôme et la saisonnalité par des indicatrices.

Au début du 20ème siècle, le développement de la publication de statistiques économiques a orienté de nombreux travaux vers la construction de méthodes de désaisonnalisation. Deux types de modèles ont été définis :

- le modèle additif : $Y_t = T_t + C_t + S_t + I_t$

- le modèle multiplicatif : $Y_t = T_t \times C_t \times S_t \times I_t$

où T_t est la tendance, C_t le cycle, S_t la saisonnalité et I_t l'irrégulier.

L'écriture du modèle multiplicatif est ici très schématique puisqu'il a en réalité été défini de diverses façons :

par exemple $Y_t = T_t \times (1 + C_t) \times (1 + S_t) \times (1 + I_t)$.

Mills, en 1924, a défini chacune des composantes de la façon suivante :

T_t : tendance : mouvement lisse, régulier, de long terme.

S_t : variations saisonnières : périodiques, avec une période annuelle (12 mois ou 4 trimestres).

C_t : variations cycliques : périodiques de façon moins marquée mais néanmoins caractérisées par un degré important de régularité.

I_t : irrégulier : le résidu.

Les années 20-30 ont vu l'émergence des grands courants de la désaisonnalisation

Dans les années 20-30, les schémas de décomposition définis plus haut étaient bien acceptés et d'autres concepts ont été fixés : l'idée que la saisonnalité varie dans le temps ; la nécessité de calculer la tendance et le cycle lorsqu'on estime la composante saisonnière ; l'impossibilité de décrire les tendances et les cycles par des formules mathématiques explicites ; la nécessité de traiter les points extrêmes. Le problème résidait surtout dans la façon de désaisonnaliser. Les travaux de

l'époque étaient inspirés par deux grandes méthodes, dont on donnera une description pour le cas d'un modèle multiplicatif.

La première, mise au point par Persons (1919), appelée méthode des "link relatives", était basée sur les médianes, pour chaque mois, des rapports entre les observations pour ce mois et celles du mois précédent. Il construisait des indices à partir de ces médianes, en déduisait des indices de tendance par lesquels il divisait les données initiales pour avoir les indices saisonniers. Il obtenait les indices saisonniers définitifs en les ajustant de façon que leur somme fasse 1. Les données cvs (corrigées de variations saisonnières) étaient les données initiales divisées par ces indices.

La seconde méthode, dite méthode des rapports à la moyenne mobile, a été introduite par Macauley (1930) et était utilisée par la Réserve Fédérale (US). Elle s'appuyait sur le calcul d'une moyenne mobile centrée d'ordre 12 pour obtenir une estimation de la tendance. Le rapport entre les données originales et cette estimation fournissait une première estimation des composantes saisonnières. Pour en éliminer l'irrégulier, on calculait les médianes (ou moyennes) de ces composantes pour chaque mois. Puis on ajustait ces nouveaux indices pour que leur somme fasse 1 et l'on obtenait ainsi les indices saisonniers définitifs. Ces deux méthodes pouvaient être adaptées au cas d'un modèle additif.

Malgré quelques critiques sur les moyennes mobiles, au moment de la découverte de l'effet Slutzky-Yule en 1927 (introduction par les moyennes mobiles de cycles artificiels dans l'irrégulier), l'idée la plus répandue dans les années 30 était bien celle qu'il ne fallait pas spécifier une forme fonctionnelle pour la tendance et la saisonnalité (par exemple un polynôme et des fonctions harmoniques ou indicatrices), et donc que les méthodes empiriques étaient préférables.

Quant au problème des valeurs extrêmes, qui agissent fortement sur les moyennes, il était parfois résolu en utilisant des médianes ou des moyennes tronquées.

D'autres types de critiques se faisaient toutefois entendre à l'égard des méthodes empiriques. Ainsi Snow (1923) estimait que la logique de la méthode de Persons n'était pas claire ; Fisher (1937) déplorait que l'on applique des méthodes empiriques ad hoc alors qu'il existait des outils mathématiques adéquats. Des recherches ont donc été faites pour élaborer des méthodes basées sur la modélisation. Celles-ci s'appuyaient en général sur une décomposition additive de la série initiale ou d'une transformation simple de cette série, et des modèles étaient posés pour la série initiale et pour chacune des composantes. Le modèle pour la série initiale était estimé à partir des données mais les modèles pour les composantes inobservables ne pouvaient être estimés qu'à condition de faire certaines hypothèses arbitraires. Les diverses méthodes différaient par le type de modèle posé et par les hypothèses faites sur les composantes.

Les développements des méthodes de désaisonnalisation ont été liés à ceux de l'informatique et des méthodes d'analyse des séries temporelles.

Après la seconde guerre mondiale, l'évolution des travaux sur la désaisonnalisation a été très liée au développement de l'informatique. Celle-ci, en accroissant la rapidité d'exécution des calculs, permit d'élaborer des méthodes plus sophistiquées qui purent être appliquées même si le nombre de séries à traiter était très important. Les méthodes les plus répandues ont été développées au bureau du Census (US). C'est là en particulier qu'a été mise au point par Julius Shiskin (1954) la méthode Census II (version informatique de la Census I) qui est à l'origine de la méthode X11 (1965). De plus, les traitements informatiques semblaient plus objectifs puisque dans certaines méthodes, des décisions faisant appel au jugement de l'utilisateur ont pu être en partie automatisées. Enfin, l'informatique a facilité l'utilisation de régressions visant à corriger les effets de jours ouvrables, c'est-à-dire les effets liés au nombre de jours travaillés dans le mois.

Les nouvelles méthodes d'analyse des séries temporelles se sont retrouvées dans les évolutions de la désaisonnalisation. L'utilisation de l'économétrie en liaison avec celle-ci s'est développée avec l'idée que la saisonnalité dans une variable économique ne peut être considérée comme un phénomène isolé, mais peut être reliée à la saisonnalité d'autres variables économiques avec laquelle elle est liée, et également que la saisonnalité elle-même peut contenir de l'information sur les relations entre les séries.

Par ailleurs, l'analyse spectrale était au départ utilisée pour chercher des périodicités exactes alors qu'on était conscient que les cycles économiques n'avaient pas une périodicité exacte : cette technique ne semblait donc pas adaptée au problème de décomposition des séries. Les progrès dans la compréhension de l'analyse spectrale ont finalement permis de l'appliquer à la désaisonnalisation. Wiener (1939,1941) et Kolmogorov (1949) puis Hannan (1967), Cleveland et Tiao (1976), Bell (1984) ont pu, par extraction de signal, estimer la composante saisonnière.

A partir de la publication des travaux de Box et Jenkins sur les modèles ARIMA (1970), et grâce aux progrès de l'informatique, l'utilisation de ces modèles s'est répandue et a été orientée vers deux types de méthodes. D'une part elle a constitué un développement important de X11 qui a évolué vers X11-ARIMA (Dagum, 1975). Dans cette nouvelle version, les modèles ARIMA sont utilisés pour prolonger la série initiale afin de limiter les révisions des estimations lorsque l'on possède un point supplémentaire. D'autre part, la modélisation ARIMA a aussi été introduite dans les méthodes de désaisonnalisation fondées sur la théorie de l'extraction du signal (Box, Hillmer et Tiao (1978), Burman (1980), Hillmer et Tiao (1983) et Hillmer, Bell et Tiao (1983)).

Les évolutions récentes

Aujourd'hui, les deux grandes philosophies de la désaisonnalisation, à savoir l'approche empirique ad hoc et l'approche par modélisation, inspirent diverses méthodes, dont certaines mêlent les deux. Parmi les méthodes actuellement les plus utilisées, on peut citer BV4 (Université technique de Berlin), Dainties (méthode officielle de la Commission Européenne), SABL (Laboratoires Bell), X11-ARIMA 88 (Statistique Canada), STL, STAMP. D'autres, mises au point plus récemment, pourraient connaître un certain succès : X12-ARIMA et TRAMO-SEATS.

Les principales critiques que l'on peut faire à chacune des deux approches sont difficiles à éviter. Ainsi, on reproche aux méthodes empiriques de ne pas s'appuyer sur la théorie statistique, ce qui rend particulièrement difficile, voire impossible, la connaissance des propriétés statistiques des estimateurs utilisés ; les méthodes basées sur les modèles sont satisfaisantes sur ce plan là mais on s'interroge sur la pertinence de la modélisation dans le cas de certaines séries très chahutées, et on invoque le manque de modèles adéquats et de théorie statistique pour les séries non stationnaires.

C'est pourquoi les améliorations que tentent d'apporter les nouvelles méthodes de désaisonnalisation ne concernent pas le principe même des méthodes existantes mais visent plutôt à corriger certains de leurs défauts. Les principales préoccupations sont tournées d'une part vers les problèmes de non symétrie des filtres en début et fin de série, et d'autre part vers l'élimination des divers effets perturbateurs qui influencent les résultats de la désaisonnalisation (points aberrants, changements de régime, effets de calendrier...).

Bilan sur les différents types de méthodes de désaisonnalisation

Schématiquement, les méthodes de désaisonnalisation peuvent être classées en deux grandes catégories : les méthodes non paramétriques et les méthodes paramétriques. Les méthodes non paramétriques, ou empiriques, permettent de décomposer la série en composantes inobservables par une procédure, souvent itérative, basée sur des lissages successifs. On peut résumer l'ensemble des lisseurs utilisés dans ces méthodes sous le nom de "régressions locales". Les régressions locales consistent à ajuster des polynômes, en général par les moindres carrés, sur des intervalles glissants (se décalant d'un point à chaque fois). Au centre de l'intervalle, la donnée lissée est la valeur, à cette date, du polynôme ajusté (la donnée lissée à la date suivante est obtenue par ajustement d'un polynôme sur l'intervalle suivant). On peut montrer que les régressions locales reviennent à appliquer des moyennes mobiles particulières lorsque les intervalles de temps sont réguliers. On peut toutefois

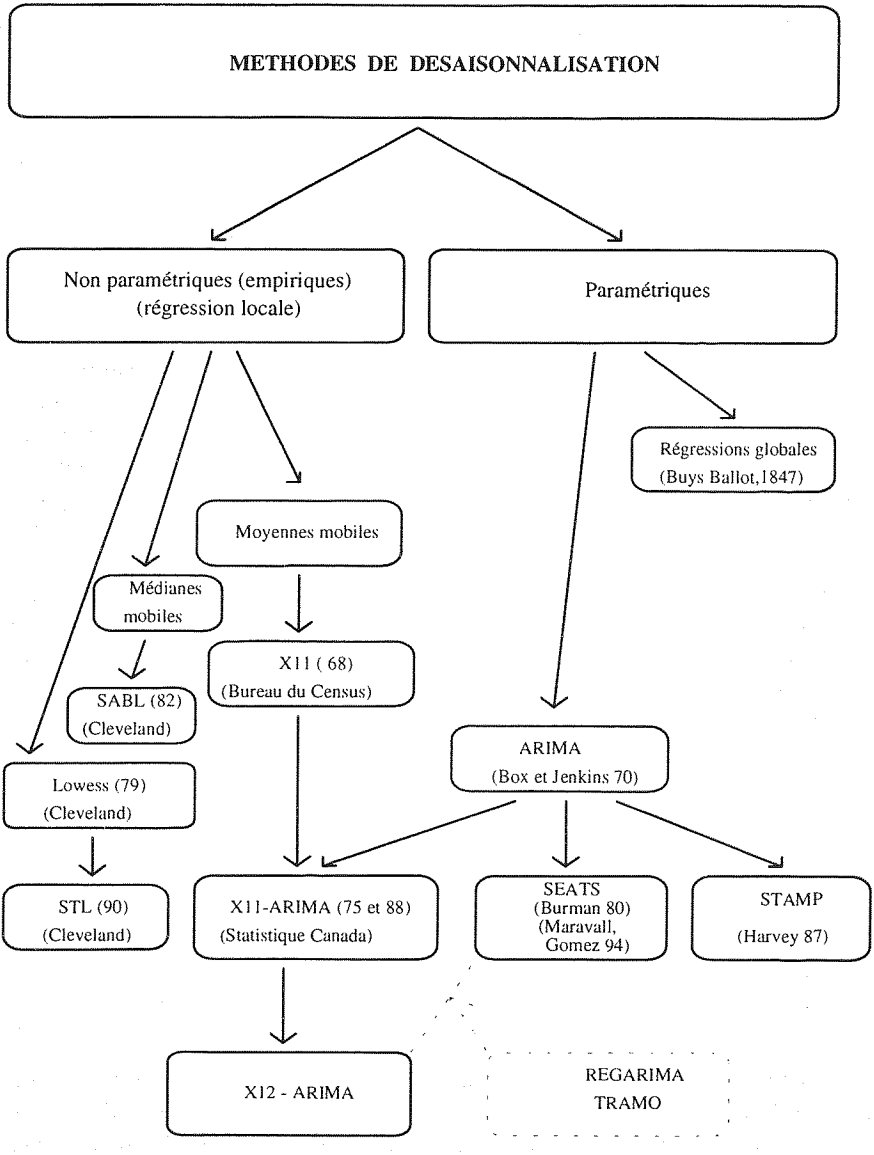
distinguer les méthodes purement basées sur les régressions locales (sans préoccupation des moyennes mobiles associées) et les méthodes utilisant directement les moyennes mobiles. Dans le premier groupe, on trouve STL (Cleveland, 1990), fondée sur le "lowess", à savoir le lissage robuste par régressions locales ; dans le second groupe, la plus célèbre est X11.

Les méthodes paramétriques peuvent elles aussi se diviser en deux grands ensembles : les méthodes par régression globale (inspirées par Buys Ballot) qui posent pour chaque composante, excepté l'irrégulier, une fonction déterministe du temps ; les méthodes basées sur des modèles stochastiques (non déterministes) : il s'agit principalement des modèles ARIMA (Box et Jenkins). Ces méthodes supposent les composantes inobservables mais modélisables par des ARIMA. Parmi celles-ci on distingue encore deux groupes : celles qui estiment les modèles des composantes à partir du modèle ARIMA de la série initiale (Burman, Bell, Hillmer, Maravall) - SEATS est la plus récente - et celles qui les estiment directement (Engle, Harvey, Todd) - par exemple la méthode STAMP -.

Les modèles ARIMA ont fait le pont entre les deux grandes approches de la désaisonnalisation. A la base même de celle-ci pour certaines méthodes par modélisation, ils ont eu à partir de 1975 un rôle important dans les méthodes par moyennes mobiles. En effet les filtres moyennes mobiles symétriques font perdre des points en début et fin de série. Ce problème est en général réglé en appliquant aux extrémités des moyennes mobiles asymétriques. Toutefois, pour réduire l'asymétrie, l'idée est venue de prolonger les séries brutes par des modèles ARIMA, ce qui a conduit au développement X11-ARIMA.

Enfin, le souci d'éliminer les effets perturbateurs qui nuisent à la qualité de la désaisonnalisation s'est manifesté dès les années 30 chez les partisans des diverses méthodes de l'époque. Puis dans les années 70-80, des auteurs ont tenté de prendre en compte les valeurs extrêmes dans la modélisation des séries (Fox (1972), Abraham et Box (1979), Denby et Martin (1979), Martin (1980), Chang (1982), Bell (1983), Hillmer, Bell et Tiao (1983)). Mais plus récemment la technique adoptée est la régression globale sur des variables particulières (de type indicatrices), qui permet de traiter toutes sortes de non-linéarités. Ce traitement préalable assez poussé des séries en amont de la désaisonnalisation constitue un deuxième pont entre les deux grandes philosophies de la désaisonnalisation : X11-ARIMA a été étendu à X12-ARIMA avec son étape RegARIMA, et SEATS peut être couplé avec TRAMO qui traite la série et la modélise par un ARIMA.

Le graphique de la page suivante résume cette présentation.


METHODES DE DESAISONNALISATION
Non paramétriques (empiriques)
(régression locale)
Paramétriques
Régressions globales
(Buys Ballot,1847)
Moyennes mobiles
Médianes
mobiles
X11 (68)
(Bureau du Census)
SABL (82)
(Cleveland)
Lowess (79)
(Cleveland)
ARIMA
(Box et Jenkins 70)
STL (90)
(Cleveland)
X11-ARIMA (75 et 88)
(Statistique Canada)
SEATS
(Burman 80)
(Maravall,
Gomez 94)
STAMP
(Harvey 87)
X12 - ARIMA
REGARIMA
TRAMO

Deux nouveaux logiciels de désaisonnalisation : X12-ARIMA et TRAMO-SEATS

Deux méthodes de désaisonnalisation ont été mises au point récemment. X12-ARIMA (1994) fait partie des méthodes empiriques : elle succède à X11-ARIMA, en y apportant quelques développements importants. SEATS (1994) (**S**ignal **E**xtraction in **A**RIMA **T**ime **S**eries), conjugué avec TRAMO (**T**ime **S**eries **R**egression with **A**RIMA Noise, **M**issing Observations, and **O**utliers), s'inspire du programme élaboré par J.P. Burman (1980) et appartient au groupe des méthodes paramétriques, basées sur une modélisation globale de la série. Il ne s'agira pas ici de faire une comparaison empirique de ces logiciels, qui pourrait constituer une étude à part entière. On en fera plutôt une présentation, qui paraît intéressante à plusieurs titres. D'abord ces deux méthodes font l'objet de diverses études réalisées par un groupe de travail à EUROSTAT, qui laissent penser qu'elles pourraient être largement utilisées dans les années à venir. Ensuite parce qu'elles illustrent bien le bilan sur les différentes approches de la désaisonnalisation présenté précédemment : en opposition quant au principe même de la décomposition d'une série, ces méthodes ont toutefois en commun le souci de capter et d'éliminer les divers effets susceptibles de perturber la désaisonnalisation, d'utiliser au mieux l'information contenue dans la série et de fournir un diagnostic détaillé des résultats. Le présent papier s'attachera principalement à décrire comment X12-ARIMA et TRAMO-SEATS tentent de répondre à ces exigences.

1. Le principe des méthodes

Une décomposition d'une série temporelle n'est pas unique. Elle dépend de la définition des composantes, qui peut être plus ou moins précise. L'utilisateur, en fonction de ses objectifs, choisira la méthode qui lui convient le mieux.

1.1. X12-ARIMA

X12 est fondée sur le même principe que X11. Elle décompose la série selon un schéma additif ou multiplicatif en tendance, saisonnalité et irrégulier. Les composantes sont obtenues à l'issue d'un processus itératif basé sur des lissages par moyennes mobiles. Dans le cas d'un schéma multiplicatif, on peut le résumer de la façon suivante. Une moyenne mobile symétrique d'ordre 12 (pour une série mensuelle, 4 pour une série trimestrielle) fournit une première estimation de la tendance puisqu'elle a pour propriétés d'éliminer les saisonnalités d'ordre 12 et de réduire le bruit. La série initiale est alors divisée par cette estimation pour donner les rapports saisonnalité-irrégulier (SI). Cette dernière série est disposée en 12 colonnes, chacune correspondant à un mois. Une moyenne mobile pondérée (d'ordre 3x3 par exemple) est ensuite appliquée à chaque colonne pour éliminer l'irrégulier et fournir

des facteurs saisonniers provisoires. Ceux-ci sont alors redisposés sous forme de série et normalisés. Les rapports SI sont divisés par ces facteurs saisonniers pour obtenir une estimation de la composante irrégulière. Les points extrêmes sont repérés sur la série de l'irrégulier : ils sont éliminés si leur écart à la moyenne est supérieur à 2,5 fois l'écart type, ou corrigés à l'aide d'une fonction de poids si leur écart à la moyenne est supérieur à 1,5 fois l'écart-type. Les étapes précédentes (sauf la première) sont réitérées en intégrant la correction des valeurs extrêmes. Une première estimation de la série désaisonnalisée est obtenue en divisant la série initiale par les nouveaux facteurs saisonniers. Une moyenne mobile de Henderson est alors appliquée à celle-ci pour une dernière estimation de la tendance. La série initiale est divisée par cette tendance et l'on fait subir aux rapports SI obtenus les étapes précédentes qui conduisent alors aux facteurs saisonniers et à l'irrégulier définitifs.

Dans X12, la définition des composantes n'est donc pas explicite. On pourrait dire brièvement que la saisonnalité doit être à peu près annulée par une moyenne mobile symétrique d'ordre égal à la périodicité de la série, et que les coefficients saisonniers sont normalisés, c'est-à-dire que leur somme sur n'importe quelle période de 12 mois doit faire approximativement 12 . Cette propriété revient à dire que la somme des valeurs de la série brute sur un an doit être environ égale à la somme des valeurs de la série corrigée des variations saisonnières (CVS) sur la même année. La tendance finale est obtenue par application sur la série CVS d'une moyenne mobile de Henderson : elle répond donc au critère de lissage correspondant à ce type de moyennes mobiles (minimisation de la somme des carrés des différences troisièmes de la série). Enfin l'irrégulier est ce qui reste après élimination de la tendance et de la saisonnalité.

1.2. SEATS

SEATS définit plus précisément les composantes, à partir de leur densité spectrale :

- la tendance présente des pics spectraux à la fréquence 0,

- la composante saisonnière présente des pics spectraux à la fréquence correspondant à la périodicité,

- la composante cyclique représente les fluctuations périodiques de période supérieure à un an ; elle présente des pics spectraux à la fréquence correspondante, entre 0 et $2\pi / s$, où s est la périodicité,

- l'irrégulier est un bruit blanc, de spectre plat.

Comme la densité spectrale d'une série peut être paramétrée par l'intermédiaire du formalisme des processus ARIMA, le principe de la décomposition par SEATS est

le suivant : le logiciel part d'une série initiale modélisée par un modèle ARIMA. Il estime ce modèle et fournit un diagnostic détaillé de l'estimation. Puis il partitionne de façon additive le spectre en spectres associés aux différentes composantes par l'application d'un filtre de type Wiener-Kolmogorov : chaque composante est alors modélisée par un ARIMA. Afin d'identifier les composantes de façon unique, il suppose qu'elles doivent vérifier la condition "canonique" qui veut que chacune des composantes (excepté l'irrégulier) soit exempte d'irrégulier. Cela signifie qu'aucun bruit additif ne peut être extrait d'une composante autre que l'irrégulier. La variance de celui-ci est donc maximisée et au contraire, la tendance, la saisonnalité et le cycle sont aussi stables que possible. Bien qu'arbitraire (puisque toute autre décomposition exprimée par la canonique plus un bruit indépendant est admissible), cela permet d'éviter la contamination des composantes par du bruit, à moins qu'il n'y ait des raisons a priori de le faire.

Sur le principe de ces méthodes, on peut émettre un rapide jugement. Par rapport à X12-ARIMA, l'approche par modélisation (SEATS) est plus confortable, dans la mesure où les propriétés de la décomposition s'appuient sur la théorie statistique. On peut ainsi connaître la variance des estimateurs et effectuer des tests. Il est de surcroît plus aisé d'obtenir des prévisions à partir de composantes modélisées. Cependant, supposer a priori une forme particulière pour les éléments de la décomposition peut être jugé trop arbitraire, d'autant plus que dans la réalité, il n'est pas certain que l'on puisse modéliser de façon satisfaisante toute série par un modèle ARIMA.

2. La modélisation ARIMA dans les logiciels de désaisonnalisation

Dans le bilan sur les différentes méthodes de désaisonnalisation, on a vu que les modèles ARIMA étaient présents dans les deux grandes familles de méthodes. Ainsi X11 a été étendue à X11-ARIMA puis X12-ARIMA. On vient par ailleurs de montrer que SEATS s'appuie également sur ce type de modélisation. L'idée de modéliser la série initiale découle de motivations différentes suivant l'approche choisie pour la décomposition. Mais dans les deux cas, les travaux de Box et Jenkins ont fait apparaître cette famille de modèles comme une technique de modélisation et d'extrapolation très puissante et bien adaptée à un traitement en masse de séries.

L'écriture générale d'un modèle ARIMA saisonnier est la suivante :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t + c$$

où : Y_t est la série chronologique étudiée

ε_t est un bruit blanc

C est une constante

S est la périodicité de la série

B est l'opérateur retard, défini par : $BX_t = X_{t-1}$

ϕ est un polynôme en B de degré p

Φ est un polynôme en B^S de degré P

θ est un polynôme en B de degré q

Θ est un polynôme en B^S de degré Q

L'écriture schématique est $(p,d,q)(P,D,Q)_s$

La modélisation ARIMA fournit des informations utiles sur la qualité des données brutes. En effet, le principe fondamental de la désaisonnalisation est l'existence d'un schéma de composition de la série. Si l'on ne peut ajuster à la série un modèle ARIMA, qui décrit la structure générale de la série en fonction de ses valeurs passées et de perturbations retardées, cela peut signifier que la série est presque purement aléatoire ou fortement perturbée par des éléments accidentels, de sorte que sa composante systématique n'est pas identifiable. On doit alors avoir un jugement critique sur la qualité des résultats de la décomposition.

2.1. Le rôle de la modélisation ARIMA dans X12-ARIMA

Le principal objectif de la modélisation ARIMA dans cette méthode est de répondre à l'exigence évoquée plus haut d'utiliser au mieux l'information contenue dans la série.

X11 apparaît comme le logiciel de référence pour la méthode de désaisonnalisation par moyennes mobiles. Le problème de la suppression de points en début et fin de série par les moyennes mobiles symétriques y est résolu par l'emploi de moyennes mobiles asymétriques, qui sont des moyennes mobiles de Musgrave. Celles-ci prolongent les moyennes mobiles de Henderson avec pour principe de minimiser les révisions des estimations des derniers points lorsque la donnée suivante sera disponible. Les autres moyennes mobiles symétriques sont prolongées par des moyennes mobiles asymétriques prédéfinies.

L'une des critiques qui étaient formulées à l'encontre de X11 était le fait que les estimations relatives aux observations les plus récentes n'étaient pas aussi fiables que les observations centrales, à cause de la nécessité d'appliquer des moyennes mobiles asymétriques en fin de série.

L'extension X11-ARIMA, développée en 1975 par Dagum, de même que X12-ARIMA, résout en partie ces problèmes : elle modélise la série initiale par un ARIMA et la prolonge d'un an au début et à la fin. Elle minimise l'erreur quadratique moyenne des estimations, et on peut montrer qu'elle minimise aussi les révisions qu'il faut apporter aux facteurs saisonniers lorsque la donnée suivante est disponible. En particulier la modélisation ARIMA est avantageuse dans les situations correspondant à des points de retournement de tendance car les modèles ARIMA captent bien les évolutions récentes.

Certes, le problème des estimations aux extrémités de séries n'est pas entièrement résolu puisqu'il faudrait bien plus d'un an supplémentaire de chaque côté pour symétriser totalement les moyennes mobiles. Cependant, l'estimation de la tendance-cycle pour la dernière observation est effectuée à l'aide d'une combinaison de coefficients des moyennes mobiles de Henderson et des coefficients de pondération du modèle ARIMA utilisés pour l'extrapolation, qui sont spécifiques à la série et peuvent donc saisir les mouvements les plus récents.

Enfin, l'addition d'une ou deux années de chiffres extrapolés, donc sans valeurs extrêmes, permet d'obtenir une meilleure estimation de l'écart-type de l'irrégulier, donc d'améliorer l'identification et le remplacement des valeurs extrêmes.

2.2. Le rôle de la modélisation ARIMA dans SEATS

Le rôle des modèles ARIMA dans SEATS est beaucoup plus rapide à définir. La décomposition de la série se fait à partir du modèle ARIMA de la série initiale et conduit à une modélisation ARIMA de chaque composante. La modélisation ARIMA est donc à la base même de la méthode, et, contrairement à X11-ARIMA, n'intervient pas simplement comme une amélioration d'une méthode qui pourrait fonctionner sans. Il faut noter que dans SEATS, les séries sont également prolongées car le filtre de Wiener-Kolmogorov fait "perdre" des points aux extrémités.

La modélisation ARIMA étant indispensable à la méthode de désaisonnalisation SEATS, il existe un modèle par défaut, le modèle Airline $(0,1,1)(0,1,1)_s$, dont l'expérience montre qu'il s'adapte bien à un grand nombre de séries, et qui sera utilisé si aucun autre n'a été spécifié.

2.3. Les procédures de modélisation ARIMA automatique dans X12-ARIMA et TRAMO

Dans X12-ARIMA et dans SEATS, l'utilisateur peut lui-même spécifier le modèle ARIMA qui lui semble le mieux adapté à la série. Mais il peut aussi avoir recours à une identification automatique du modèle. Ainsi, le logiciel TRAMO, qui peut être aisément couplé avec SEATS, propose une telle procédure. Dans X12-ARIMA, elle est directement intégrée dans le logiciel. Toutefois les procédures diffèrent dans les deux méthodes.

Dans TRAMO, l'identification automatique du modèle ARIMA se fait en deux temps. La première étape consiste en la détermination des facteurs de différenciation $\left(1-B\right)\left(1-B^{s}\right)$ et de la constante. Elle est réalisée par itération sur la séquence de modèles AR et ARMA (1,1), qui ont une structure multiplicative lorsque les données sont saisonnières. Cette procédure est basée sur les résultats de Tiao et Tsay (1983) et Tsay (1984). Les différences régulières et saisonnières sont d'ordre maximum respectivement 2 et 1.

Dans la deuxième étape, le logiciel procède à l'identification du modèle ARMA pour la série stationnaire. Elle se fait suivant la procédure de Hannan-Rissanen, avec quelques modifications. La recherche consiste à balayer les valeurs suivantes :

$0 \leq (p,q) \leq 3$ et $0 \leq (P,Q) \leq 2$

Elle est faite séquentiellement (pour des polynômes réguliers fixés, on obtient les polynômes saisonniers et vice-versa) et les ordres finals des polynômes sont choisis selon le critère BIC (Critère d'Information Bayesien), avec une contrainte possible au nom du principe de parcimonie et en faveur des modèles équilibrés (mêmes ordres AR et MA). Le critère BIC mesure l'écart entre la vraie loi des observations et le modèle proposé.

Dans X12-ARIMA, la modélisation automatique se fait suivant le même principe que dans X11-ARIMA : dans le programme sont prédéfinis cinq modèles saisonniers, qui ont été choisis parce qu'ils modélisaient de façon satisfaisante un grand nombre de séries aussi bien en termes d'ajustement sur le passé qu'en termes de projections pour les trois dernières années. Ces modèles sont testés et le meilleur est conservé. Il s'agit des modèles dont les parties non saisonnières sont (0,1,1), (0,1,2), (2,1,0), (0,2,2) et (2,1,2) et avec toujours la même partie saisonnière $(0,1,1)_S$. L'utilisateur a la possibilité d'introduire un ensemble de modèles qu'il souhaite tester automatiquement. Si l'utilisateur entre un nombre important de modèles, cela peut revenir à effectuer un balayage des paramètres, comme dans SEATS.

Les critères de qualité de la modélisation sont les suivants : la somme des carrés des erreurs de prévision ; le test du "portmanteau" (Ljung-Box) ; les prévisions "hors échantillon" ; si plusieurs modèles incorporés à l'option automatisée satisfont aux critères d'acceptation, le programme utilise celui qui donne les meilleurs résultats pour une année d'extrapolations rétrospectives.

3. Le traitement préalable des séries : couplage régression-ARIMA

3.1. Définition du modèle Reg-ARIMA

Les modèles ARIMA sont des modèles linéaires. Lorsqu'on les utilise pour modéliser une série, on suppose donc que la série est exempte de non-linéarités. Or dans la réalité, cette hypothèse n'est pas vérifiée pour un grand nombre de séries qui peuvent présenter des points aberrants, des changements brutaux de niveau, des effets de calendrier... Afin que la modélisation ARIMA soit pertinente, il convient donc de les éliminer préalablement.

Le traitement des non-linéarités dans les méthodes de désaisonnalisation se justifie à plusieurs titres. Le programme SEATS suppose explicitement que la série initiale en est exempte et n'effectue lui-même aucune correction. Dans X11-ARIMA, il existe bien des traitements pour certains types d'effets perturbateurs tels que les points aberrants, les effets de jours ouvrables et les effets de jours fériés mobiles. Mais, bien que la première étape du programme permette de réaliser quelques ajustements préalables, ils sont pour la plupart effectués au cours du processus de décomposition de la série. Or des tests empiriques tendent à prouver que l'on obtient de meilleurs résultats pour la désaisonnalisation lorsqu'ils sont effectués en amont du processus.

La solution proposée dans les deux cas est l'introduction d'une régression linéaire destinée à intégrer l'information que l'on possède sur la série et à capter les perturbations identifiables. Ce sont les résidus de cette régression qui seront alors modélisés par un ARIMA. Ce couplage régression-ARIMA constitue l'apport principal de X12-ARIMA par rapport à X11-ARIMA et de l'utilisation de TRAMO avec SEATS.

Le modèle Reg-ARIMA est défini de la même façon dans TRAMO et dans X12-ARIMA :

Si Y_t est la série initiale, alors le programme ajuste le modèle de régression suivant :

$$Y_t = \sum \beta_i X_{it} + z_t$$

où les β_i sont les coefficients de la régression, les X_{it} les variables explicatives qui sont soit fournies par l'utilisateur, soit prédéfinies par le programme (voir des exemples plus loin), et les z_t les résidus.

Il modélise ensuite les résidus de cette régression par un ARIMA :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D z_t = \theta(B)\Theta(B^s)a_t + c$$

où a_t est un bruit blanc (i.i.d $(0,\sigma^2)$, et gaussien dans le modèle de TRAMO).

La modélisation de z_t par un ARIMA met en lumière le fait que les résidus d'une régression dans le cas de séries temporelles sont la plupart du temps autocorrélés ; faire l'hypothèse qu'ils ne le sont pas conduirait à des résultats inexacts.

La dénomination Reg-ARIMA est la combinaison de ces deux modèles :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D(Y_t - \sum \beta_i X_{it}) = \theta(B)\Theta(B^s)a_t + c$$

Ce modèle s'interprète comme une généralisation d'un modèle de régression pur, où les résidus sont supposés être un bruit blanc, à un modèle de régression où les résidus suivent un processus ARIMA.

Les logiciels TRAMO et X12-ARIMA offrent à l'utilisateur la possibilité de choisir ses variables explicatives, s'il possède des informations sur la série qu'il cherche à modéliser, ou bien d'utiliser celles qui sont prédéfinies par le programme. Par exemple on peut introduire des variables exogènes dont on sait qu'elles sont liées à la variable étudiée. Les variables prédéfinies sont destinées à isoler des perturbations ou des effets interprétables :

- les points aberrants ("additive outliers" et "innovationnal outliers"),
- les ruptures de niveau ("level shifts"),
- les changements transitoires de niveau ("temporary changes"),
- les saisonnalités fixes,
- les changements de régime (modification de la saisonnalité),
- les effets de jours fériés,
- les effets de jours ouvrables.

Tous ces effets sont modélisés par des variables particulières (indicatrices par exemple) mais la méthodologie diffère parfois entre les deux logiciels.

3.2. Le traitement des points aberrants et des changements de niveau (par ruptures ou temporaires) :

- les variables :

TRAMO propose différentes variables d'intervention du type :

- 1. variables indicatrices (additive outliers),

- 2. toutes les séquences possibles de 1 et de 0,

- 3. $1/(1-\delta B)$ appliqué à toute séquence de 1 et de 0, où $0 < \delta \leq 1$,

- 4. $1/(1-\delta_s B^s)$ appliqué à toute séquence de 1 et de 0, où $0 < \delta_s \leq 1$

- 5. $1/(1-B)(1-B^S)$ appliqué à toute séquence de 1 et de 0.

Interprétation : Les variables indicatrices et les séquences de 1 et de 0 captent des perturbations ponctuelles ou répétitives. Le troisième type de variables peut s'écrire de la façon suivante :

si X_t est une séquence de 1 et de 0, alors

$$(1-\delta B)X_t = \left(\sum_{i=0}^{+\infty} \delta^i B^i\right) X_t = \sum_{i=0}^{+\infty} \delta^i X_{t-i} \cdot$$

Ainsi, à chaque instant t, la valeur de cette variable est égale à la valeur de X_t (soit 0 ou 1) plus les répercussions des valeurs 1 (perturbations) antérieures de X affectées d'un coefficient d'autant plus faible que ces perturbations sont éloignées dans le temps. Par exemple, si $X_t = 1$ et si les seules dates antérieures pour lesquelles X vaut 1 sont (t-i) et (t-j) alors $(1-\delta B)X_t = 1 + \delta^{t-i} + \delta^{t-j}$.

L'interprétation du quatrième type de variables est quasiment la même, sauf qu'à un instant t, seules les perturbations distantes de t d'un multiple de la période s sont répercutées.

Enfin, la dernière variable représente, à chaque instant t, le produit du nombre de perturbations précédentes et du nombre de perturbations précédentes distantes de t d'un multiple de la période.

X12-ARIMA modélise les points aberrants par des variables indicatrices, les ruptures de niveau et les changements transitoires de niveau par les variables suivantes :

- si une rupture de niveau survient à un instant t_0, la variable

$$LS_t^{t_0} = \begin{cases} -1 \text{ pour } t < t_0 \\ 0 \text{ pour } t \geq t_0 \end{cases}$$

- si un changement transitoire de niveau a lieu entre t_0 et t_1, la variable

$$TR_t^{(t_0,t_1)} = \begin{cases} -1 \text{ pour } t \leq t_0 \\ (t-t_0)/(t_1-t_0) - 1 \text{ pour } t_0 < t < t_1 \\ 0 \text{ pour } t \geq t_1 \end{cases}$$

- la méthodologie :

S'ils ne sont pas spécifiés, les points aberrants et autres effets peuvent être détectés automatiquement par le programme. TRAMO et X12 utilisent pour cela des méthodologies différentes.

TRAMO a une approche similaire à celle de Chen et Liu (1993), avec quelques modifications. Les paramètres de la régression sont initialisés par les moindres carrés ordinaires et les paramètres du modèle ARMA sont alors estimés avec deux régressions, comme dans Hannan et Rissanen (1982). Puis le filtre de Kalman fournit la série des résidus, et une nouvelle estimation des paramètres de régression est obtenue. Pour chaque observation, des tests de Student sont effectués, pour chaque type de non linéarité comme dans Chen et Liu (1993). Les points aberrants sont modifiés un par un et à chaque fois, une nouvelle estimation des paramètres du modèle est obtenue. Une fois cette première séquence achevée, une régression multiple est effectuée et si des points aberrants sont détectés, le programme revient à la première séquence et itère jusqu'à ce qu'aucun point aberrant ne soit éliminé dans la régression multiple.

Dans X12-ARIMA, la détection des points aberrants est basée sur la méthodologie de Chang et Tiao (1983) avec des extensions et des modifications (Bell, 1983,1994 et Otto et Bell ,1990). L'approche générale est similaire à la régression stepwise (GLS), où les variables candidates sont les variables précédemment décrites pour toutes les dates où la détection de points aberrants est effectuée. Le programme calcule la statistique de Student pour la significativité de chaque type de non-linéarité à chaque date, cherche celles qui sont significatives et ajoute les variables correspondantes au modèle. Les valeurs critiques des tests peuvent être définies par

l'utilisateur. Certains tests sont effectués pour repérer si les ruptures de niveau sont provisoires, ce qui est important pour la prévision.

3.3. Le traitement des saisonnalités fixes et des effets de calendrier

- les saisonnalités fixes

Dans TRAMO, des variables de type indicatrices peuvent permettre de prendre en compte ces effets.

X12-ARIMA propose deux types de variables périodiques : soit des variables indicatrices, soit des variables de Fourier (trigonométriques).

- les effets de jours fériés

Ces effets ont lieu lorsqu'il s'agit de jour fériés qui ne tombent par toujours à la même date et peuvent survenir à l'un ou l'autre de deux mois (ou trimestres) consécutifs ; ils ne sont donc pas pris en compte dans la correction des variations saisonnières.

Dans les deux logiciels, des variables sont construites pour évaluer l'« effet Pâques ». X12-ARIMA traite aussi l'effet « Thanksgiving » et l'effet « Labor day ».

Le modèle utilisé par X12-ARIMA suppose un changement constant du niveau de l'activité journalière durant un nombre spécifié de jours avant le jour férié. Pour le Thanksgiving, il suppose un changement constant de l'activité journalière à partir d'un certain nombre de jours avant et jusqu'à un certain nombre de jours après le Thanksgiving.

- la correction des effets de jours ouvrables

TRAMO corrige ces effets en introduisant des variables spécifiques dans la partie régressive de la modélisation REGARIMA. En revanche, SEATS ne semble pas réaliser de traitements relatifs aux jours ouvrables.

Au contraire, dans X12-ARIMA, ils peuvent non seulement être pris en compte par des variables spécifiques dans la régression, mais sont à nouveau traités dans le processus itératif de décomposition de la série.

Dans l'étape REGARIMA, une option qui repose sur le critère d'Akaike (AIC) permet de déterminer si des variables de jours ouvrables doivent être introduites dans la régression. Plusieurs sortes de variables sont prédéfinies dans le programme pour modéliser les effets de jours ouvrables, en fonction du type de la série. Celle-ci peut en effet être de type flux (les données mensuelles sont des sommes de données journalières) ou de type stock (valeur à un jour donné du mois).

Dans certaines étapes de la décomposition proprement dite, le logiciel effectue un nouveau traitement des effets de jours ouvrables par régression de l'irrégulier sur des variables particulières. Cependant, les tests de Fisher pour ce type de régression ne sont pas toujours très fiables dans la mesure où l'irrégulier présente souvent de l'autocorrélation et de l'hétéroscédasticité. C'est pourquoi la modélisation des effets de jours ouvrables dans l'étape RegARIMA (sur la série initiale) semble préférable, sauf lorsque la tendance est particulièrement chahutée.

3.4. Autres traitements

Traitement des observations manquantes

Les deux logiciels offrent la possibilité de remplacer les observations manquantes. Pour cela, une valeur "bizarre" (par exemple -9999) est imputée aux dates où les observations sont manquantes et ces observations sont alors traitées comme des valeurs extrêmes, par introduction des indicatrices correspondantes. La valeur de remplacement est la différence entre la valeur imputée et le paramètre de régression estimé pour l'indicatrice. En réalité, il existe quelques contraintes : dans X12-ARIMA, le nombre d'observations manquantes ne doit pas être trop important ; dans TRAMO, lorsque certaines valeurs manquantes initiales ne sont pas estimables (paramètres libres), elles sont remplacées par des valeurs "bizarres" pour toute la suite des traitements et les estimations affectées par ces valeurs sont alors facilement repérables.

TRAMO propose une autre méthodologie de traitement des observations manquantes, basée sur l'espérance des valeurs manquantes conditionnellement aux données disponibles. Le principe est décrit dans Gomez et Maravall (1994).

Traitement des changements de régime, des effets "promotions"

Il peut arriver que la saisonnalité d'une série se modifie à partir d'une certaine date : on parle alors de "changement de régime". Dans X12-ARIMA, ce phénomène est traité en introduisant des variables de saisonnalité fixe sur chacune des deux périodes. Pour cela, le moment où se produit la rupture doit être connu de l'utilisateur.

Parfois, la série peut être l'objet d'une forte variation que l'on sait interpréter, et qui a lieu sur une période non habituelle (sinon le phénomène serait traité par l'ajustement saisonnier). Il peut s'agir par exemple de promotions pour écouler des stocks. X12-ARIMA propose un traitement de cet effet, à condition là encore que l'utilisateur détermine précisément la période en question.

4. Diagnostics

Les programmes X12-ARIMA et TRAMO-SEATS fournissent à chaque étape diverses statistiques que l'on ne présentera pas ici dans leur ensemble. On exposera quelques diagnostics relatifs à la qualité de la désaisonnalisation, qui constituent des nouveautés par rapport à des logiciels plus anciens.

Le premier élément de diagnostic est l'analyse du spectre de l'irrégulier. Celle-ci permet de détecter la présence de saisonnalité résiduelle et de tester la présence d'effets de jours ouvrables. Il se peut en effet que, dans une série déjà corrigée des variations saisonnières et des effets de jours ouvrables, ces éléments soient encore présents. Dans le cas d'une application directe d'un des logiciels sur la série, cela peut être dû à une mauvaise utilisation des procédures. Mais souvent, cette situation se produit lorsque la série n'a pas été traitée directement mais est l'agrégation d'un ensemble de séries qui, elles, ont été ajustées par les logiciels. En effet, dans certaines "sous-séries", les effets saisonniers et les effets de jours ouvrables sont parfois difficiles à détecter et leur correction n'est pas satisfaisante, ce qui peut créer des effets résiduels dans la série agrégée.

Le second élément important du diagnostic porte sur la stabilité de l'ajustement. En effet, on peut espérer que lorsqu'une nouvelle donnée est disponible, les estimations obtenues en intégrant cette nouvelle donnée ne diffèrent pas trop des estimations précédentes. A moins qu'il n'existe réellement une très grande variabilité dans la saisonnalité ou dans la tendance, des révisions importantes sont plutôt le signe d'une mauvaise qualité d'ajustement. Le principe du diagnostic de stabilité n'est pas nouveau mais la façon dont il est mis en œuvre dans X12-ARIMA est une innovation par rapport à X11-ARIMA. X12-ARIMA propose deux types de diagnostics de stabilité. Le premier consiste à faire tourner le programme sur au moins quatre subdivisions de la série ("sliding spans"). Pour un mois commun à deux intervalles, on compare les résultats obtenus sur les deux intervalles, aussi bien pour la série ajustée que pour les évolutions d'un mois à un autre. Le second type de diagnostics considère les révisions historiques : pour une date donnée, le programme analyse les différences entre les résultats obtenus par ajustement sur la période antérieure à cette date (c'est-à-dire la période pour laquelle cette date était la dernière de la série) et ceux obtenus sur la totalité de la série. Les comparaisons portent sur la série ajustée, les évolutions d'un mois à un autre, la tendance, les évolutions de la tendance d'un mois à un autre.

D'une façon plus générale, on attire l'attention sur le fait que les deux logiciels produisent de nombreuses statistiques et divers graphiques qui fournissent à l'utilisateur une information précieuse sur la qualité de l'ajustement et permettent de le guider pour apporter des améliorations si elle n'est pas satisfaisante.

BIBLIOGRAPHIE

Bell W. R., Chen B., Findley D. F., Monsell B. C., Otto M. C. (1996) : New capabilities and methods of the X-12-ARIMA seasonal adjustment program.

Bell W. R., Hillmer S.C. (1992) : « Issues involved with the seasonal adjustment of economic times series », *Modelling seasonality,*83-138.

Bureau Of The Census (1996) : *X12-ARIMA Reference Manual* (Beta Version).

Cleveland W. S. (1983) : *Seasonal and calendar adjustment*, Handbook of Statistics, 3, 39-92.

Droesbeke J. J., Tassi P. (1990) : *Histoire de la statistique*, Que sais-je? PUF, Paris

Dagum E. B. (1980) : *La méthode de désaisonnalisation X-11-ARMMI*, Statistique Canada.

Fischer B. (1995) : *Decomposition of times series : comparing different methods in theory and in practice*, EUROSTAT.

Gomez V., Maravall A. (1996) : *Programs TRAMO and SEATS Instructions for the User* (Beta Version).

Hylleberg S. (1992). « The historical perspective », *Modelling seasonality,*15-25.

AJUSTEMENT DES SERIES SAISONNIERES : METHODES AD HOC CONTRE METHODES D'EXTRACTION DE SIGNAUX

Christophe Planas

1 Introduction

La production totale de l'industrie française (PTIF) (sauf construction) présentée sur la *figure 1* (en annexe) constitue une série chronologique mensuelle qui s'étend de janvier 1985 à avril 1996. Cette série est caractérisée par un important comportement saisonnier. Une pratique commune des analystes consiste à supprimer les mouvements saisonniers afin de faciliter l'interprétation. Les taux de croissance de la production, en particulier, sont l'objet d'une attention minutieuse. A EUROSTAT, deux programmes sont utilisés pour l'ajustement des séries saisonnières : X12-ARIMA du Bureau du Census, dernière version de la famille X11, et SEATS-TRAMO (voir Gomez et Maravall, 1996), qui met en oeuvre l'approche basée sur les modèles autorégressifs intégrés à moyenne mobile (ARIMA). Cette décomposition des séries chronologiques, dénoté AMB pour "ARIMA-model-based", a été dévelopée par Burman (1980), et Box, Hillmer et Tiao (1978) entre autres. En supprimant les variations saisonnières de la série PTIF avec les deux programmes, il apparut que les fluctuations de court-terme de la série ajustée étaient plus erratiques avec X12 qu'avec l'approche AMB. Cette plus grande variabilité était retrouvée dans les taux de croissance mensuels de la série ajustée.

Dans cet article, nous présentons de façon générale les filtres centraux d'ajustement saisonnier de X12 ainsi que le filtre de Wiener-Kolmogorov impliqué dans l'approche AMB. Explorant les propriétés des différents filtres dans le domaine des fréquences, certains traits particuliers liés aux mouvements de court-terme sont mis en évidence. En particulier, il est montré que certains filtres d'ajustement saisonniers contenus dans X11 surestiment les mouvements de court-terme des séries. L'argument de l'analyse concerne tous les filtres d'ajustement à l'exception du filtre impliquant une moyenne mobile saisonnière 3x3, et ceci quelquesoit la longueur du filtre de Henderson choisie. Les conséquences pratiques sont alors vérifiées sur la série PTIF. Nous présentons maintenant le cadre théorique général.

2 Cadre d'analyse général

Nous nous intéressons au problème de la décomposition d'une série chronologique x_t en somme de composantes inobservées, une saisonnière s_t et une nonsaisonnière n_t, d'après la relation :

$$x_t = s_t + n_t , \tag{2.1}$$

où les deux composantes sont supposées indépendantes. Aux variables x_t, s_t et n_t sont associés les spectres $g_x(w)$, $g_s(w)$ et $g_n(w)$, où w représente une fréquence telle que $w \in [0,\pi]$. La partie nonsaisonnière de la série est souvent perçue comme la somme d'une tendance et d'une composante irrégulière. Des modèles déterministes incluant des variables binaires ou des fonctions trigonométriques ont été utilisés pour décrire les fluctuations saisonnières des séries temporelles (voir Hylleberg 1986). La reconnaissance graduelle que les fluctuations saisonnières évoluent dans le temps et ne suivent pas des chemins fixes a conduit à des procédures plus flexibles. Celles-ci ont été développées sur la base des méthodes dites à moyenne mobile. Les filtres à moyenne mobile les plus souvent utilisés pour estimer les composantes ont en commun la propriété de linéarité et, pour les observations pas trop proches de la fin de l'échantillon, de symétrie. Si B désigne l'opérateur retard tel que $Bx_t = x_{t-1}$, alors un filtre linéaire à moyenne mobile peut être représenté comme :

$$\hat{s} = a(B)x_t$$

$$= \left[a_0 + \sum_{k=1}^{r} a_k \left(B^k + F^k \right) \right] x_t \tag{2.2}$$

Les filtres symétriques sont considérés parce qu'ils n'impliquent aucun effet de phase. L'absence d'effet de phase est important : il serait particulièrement génant pour les analystes de travailler sur des séries dont les figures initiales et ajustées présentent des ondulations désynchronisées.

Les différences entre les méthodes à moyenne mobile proviennent principalement de la façon dont le filtre est construit. Dans cet article, nous nous intéressons à deux types de filtres : les filtres ad hoc, c'est-à-dire les filtres de X11 et des versions successives, et le filtre de Wiener-Kolmogorov tel que rendu disponible par la théorie de l'extraction des signaux et utilisé dans l'approche AMB pour décomposer les séries temporelles. Afin d'analyser ces types de filtre, il est nécessaire d'introduire quelques concepts classiques. Les composantes inobservées étant construites de sorte à capturer les variations d'une série à certaines fréquences, il convient de conduire l'analyse des filtres à moyenne mobile dans le domaine des fréquences. La fonction de gain s'avère alors utile à l'analyse. Si l'on prend la

transformée de Fourier $B = e^{-iw}$, alors la fonction de réponse dans le domaine fréquenciel associée avec $a(B)$ est donnée par :

$$a(w) = \sum_{k=-r}^{r} a_k e^{-ikw} .$$

Le gain du filtre est directement obtenu comme $|a(w)|^2$. Le gain d'un filtre mesure l'étendue dans laquelle la contribution d'une fréquence w à la variance de la série est transmise à l'estimateur de la composante. C'est la fonction qui relie le spectre de la série input au spectre de l'estimateur de la composante $g_{\hat{s}}(w)$ d'après

$$g_{\hat{s}}(w) = |a(w)|^2 g_x(w) .$$

Spectre de l'estimateur et fonction de gain seront les deux principaux instruments de l'analyse.

3 Filtres ad hoc : les filtres d'ajustement de X11

Les décompositions de séries chronologiques dans X11 peuvent être multiplicative, additive ou log-additive. Dans ces deux derniers cas, les filtres utilisés sont linéaires et identiques, puisqu'ils s'appliquent sur la série brute ou sur sa transformation logarithmique. Les décompositions multiplicatives diffèrent sous certains aspects. Cependant, Young (1968) a montré que les filtres linéaires peuvent être vus comme des approximations de l'approche multiplicative. Seules certaines nonlinéarités sont occultées par l'approximation linéaire, et selon Young ces nonlinéarités ne sont pas très importantes en général. Dans le cas de la série PTIF, la décomposition conduite est de type additif.

Nous nous concentrons sur les filtres historiques ; il y a plusieurs raisons pour cela. D'abord, les dernières versions de X11 telles que X12-ARIMA contiennent toujours les filtres centraux de X11. Ensuite, vers la fin des échantillons, X12-ARIMA utilise les prévisions des observations futures plutôt qu'un filtre asymétrique, pourvus que ces prévisions soient jugées suffisamment précises. Enfin, pour l'application présentée, le même modèle a été considéré en X12-ARIMA et en SEATS pour former ces prévisions. Il sera tout de même vérifié que la troncation éventuelle du filtre n'affecte pas la pertinence des résultats.

Les filtres linéaires de X11 peuvent être vus comme des compositions de moyennes mobiles. Une explication détaillée des procédures en oeuvre peut être trouvée dans Wallis (1974, 1982). Bell et Monsell (1992) ont beaucoup simplifié la reconstruction des filtres additifs de X11 : ils donnent explicitement les expressions

des filtres impliqués dans chaque composition. Selon le filtre choisi à chaque étape, un résultat différent est obtenu. Les options disponibles pour les moyennes mobiles saisonnières sont 3x3, 3x3 suivi par 3x5 (défaut), 3x5, 3x9 et moyenne mobile saisonnière à trois termes. Ces filtres sont combinés avec un filtre de lissage de Henderson, dont la longueur standard est de 9, 13, ou 23 termes. Les présentations graphiques des filtres et des gains associés sont disponibles dans de nombreux articles, mais la présentation la plus complète est celle de Bell et Monsell (1992).

Par commodité, les gains des filtres d'ajustement sont reproduits. En particulier, les gains des filtres d'ajustement de défaut, 3x3, 3x9, 3 termes, associés à un filtre de Henderson de 13 termes sont présentés. Le filtre 3x5 est omis parce que son gain n'est pas très différent de celui du défaut (voir Bell et Monsell). En pratique, le filtre de Henderson de 13 termes est le plus utilisé. D'autres longueurs auraient pu être considérées : cela aurait affecté les gains dans la région des basses fréquences et aussi entre les fréquences $\pi/6$ et $\pi/5$, tout en laissant quasi intacts les gains pour les fréquences situées au-delà de $\pi/5$.

Les graphiques présentés sur la *figure 2* (en annexe) illustrent comment le filtrage fonctionne dans le domaine des fréquences. La composante saisonnière est construite de sorte à capturer les mouvements de la série se produisant avec une fréquence saisonnière. Le filtre d'ajustement doit donc annihiler la variabilité associée aux fréquences saisonnières, et laisser inchangées les variations aux autres fréquences. En accord avec ceci, les gains des filtres d'ajustement de X11 présentés sur la figure 2 montrent cette structure dite de "bandpass" : ils ont un gain de 0 autour des fréquences saisonnières et un gain proche de 1 dans les autres régions. L'amplitude des bandes de fréquences où le gain est nul est liée à la stabilité des fluctuations saisonnières qui sont supposées être effacées : le filtre 3x9 correspond à une saisonnalité stable tandis que le filtre de 3 termes serait adéquat pour une saisonnalité relativement instable. Pour les séries dont les caractéristiques seraient plus accentuées que ces deux possibilités, alors respectivement trop ou pas assez de mouvements saisonniers seraient gommés par la simple application de ces filtres (voir par exemple Fiorentini et Planas 1996).

Mais le point le plus important qui apparaît sur la *figure 2* (en annexe) est que, pour trois des quatre filtres, le gain des filtres d'ajustement centraux de X11 est plus grand que 1 pour des fréquences comprises entre la fréquence saisonnière fondamentale et la fréquence de Nyquist. Le seul filtre qui ne partage pas cette caractéristique est le filtre d'ajustement 3x3. L'utilisation d'autres longueurs pour le filtre de Henderson n'affecte pas cette propriété des filtres d'ajustement de X11.

4 Extraction de signaux dans les modèles ARIMA avec le filtre de Wiener-Kolmogorov

La construction du filtre de Wiener-Kolmogorov (WK) est exposée dans Whittle (1963). Sous les hypothèses que les composantes sont indépendantes et qu'une réalisation infinie de la série est disponible, the filtre de WK est obtenu comme le ratio du spectre de la composante au spectre de la série. Pour les séries stationnaires, l'estimateur de la saisonnalité est donné par :

$$\hat{s}_t = \nu_s(B)x_t\,, \tag{4.1}$$

où, si l'on utilise la transformée de Fourier $B = e^{-iw}$,

$$\nu_s(w) = \frac{g_s(w)}{g_x(w)}. \tag{4.2}$$

Dans le domaine temporel, le filtre de WK correspond au ratio de la fonction génératrice des autocorrélations (ACGF) de la composante à l'ACGF de la série. Le filtre de WK est bâti de sorte à minimizer la moyenne du carré des erreurs sur l'estimateur. Il produit donc la projection linéaire de la composante sur la série observée. De plus, il a été montré que le filtre de WK donne des estimateurs optimaux y compris dans le cas de séries nonstationnaires (voir, par exemple, Pierce (1979) ou Bell (1984)).

Cleveland et Tiao (1976), Burman (1980), ont suggéré d'utiliser la théorie de l'extraction de signaux en conjonction avec la spécification de modèles stochastiques linéaires de type ARIMApour la série et ses composantes. Une raison pour cela est que les modèles ARIMA permettent une paramétrisation très simple d'un spectre. Par conséquent, il est supposé que les composantes suivent des processus

$$\phi_s(B)s_t = \theta_s(B)a_{st}$$

$$\phi_n(B)n_t = \theta_n(B)a_{nt} \tag{4.3}$$

où $\phi_\bullet(B)$ et $\theta_\bullet(B)$ représentent des polynômes finis en B dont les racines se situent sur ou en dehors du cercle unitaire. Les variables a_{st} et a_{nt} sont des bruits blancs indépendants de variance V_s et V_n. Les polynômes $\phi_s(B)$ et $\phi_n(B)$ n'ont aucune racine en commun, tandis que les polynômes $\theta_s(B)$ et $\theta_n(B)$ sont supposés ne pas avoir de racine unitaire en commun.

Les équations (4.3) impliquent que la série observée x_t suit un modèle ARIMA du type :

$$\phi(B)x_t = \theta(B)a_t, \tag{4.4}$$

où a_t est un bruit blanc. Le polynôme AR $\phi(B)$ vérifie $\phi(B) = \phi_s(B)\,\phi_n(B)$. En pratique, l'identification et l'estimation du modèle (4.4) est conduit selon les techniques bien connues de Box-Jenkins (voir Box et Jenkins, 1970). La répartition des racines de $\phi(B)$ entre les polynômes $\phi_s(B)$ et $\phi_n(B)$ est accomplie selon les profils que les composantes sont supposées montrer. Par exemple, les racines complexes conjuguées dont la période est plus petite que l'année sont assignées à la composante saisonnière. Pour les polynômes MA et les variances des innovations V_n et V_s, une telle procédure d'identification n'est pas disponible. En général, si une série temporelle admet une décomposition en composantes inobservées, le nombre de décompositions admissibles est infini. Dans l'approche AMB, la sélection d'une décomposition particulière est conduite de sorte à maximiser la variance de la composante irrégulière (voir Hillmer and Tiao 1982). Une décomposition dite canonique est ainsi obtenue, les composantes canoniques présentant un zéro dans leur spectre.

Dans le domaine temporel et après simplification, l'estimateur optimal de la composante saisonnière est donné par :

$$\hat{s}_t = V_s \frac{\theta_s(B)\theta_s(F)\phi_n(B)\phi_n(F)}{\theta(B)\theta(F)} x_t$$

Le filtre historique de WK est donc symétrique, tout comme les filtres historiques de X11. Il est convergent, et donc valide pour calculer les estimateurs des signaux aux instants centraux de l'échantillon. Près des bords des échantillons, les observations futures sont remplacées par leur prévision obtenues à partir de (4.4), et les estimateurs préliminaires sont ainsi dérivés.

L'analyse du filtre de WK dans le domaine des fréquences montre que, comme les filtres de X11, il présente une structure de "band-pass", littéralement de "bande de passage". D'après (4.1), et en utilisant $g_x(w) = g_n(w) + g_s(w)$, on peut écrire

$$\nu_s(w) = \frac{1}{1 + \dfrac{g_n(w)}{g_s(w)}}. \tag{4.5}$$

Quand la contribution relative de la composante saisonnière est large à une fréquence w^*, on a $g_n(w^*)/g_s(w^*) \simeq 0$. Dans ce cas, une grande partie du spectre

de la série observée est utilisée pour l'estimation du signal : le gain du filtre à cette fréquence sera proche de 1. Inversement, si la contribution relative est faible à une fréquence particulière, le filtre de WK l'ignore simplement lors de l'estimation de la composante. Par exemple, si la composante nonsaisonnière n_t contient une tendance nonstationnaire, le spectre de n_t est infini dans la région des basses fréquences, et nous aurons $g_n(w^*)/g_s(w^*) \to \infty$. Il s'ensuit que $\nu_s(w^*) \simeq 0$: le gain sera proche de zéro dans cette région, et aucune variation de basse fréquence ne sera transmise à la composante saisonnière. Etant donné que tout deux sont des moyennes mobiles, l'interprétation des filtres de X11 et du filtre de WK dans le domaine des fréquences est similaire. La différence majeure réside dans le fait que le filtre de WK s'adapte de lui-même à la série analysée tandis que les filtres de X11 ne dépendent pas des propriétés des séries : ils sont ad hoc (voir par exemple Maravall (1993b)).

Une autre conséquence immédiate de (4.5) est que le gain du filtre de WK ne peut jamais être plus grand que 1 : la variabilité de la série à chaque fréquence est partagée entre les composantes, mais elle ne peut jamais être amplifiée. Cette différence avec les filtres empiriques de X11 a plusieurs implications. Sauf pour le filtre d'ajustement saisonnier 3x3, les filtres d'ajustement de X11 n'ont pas d'interprétation en termes de projection linéaire : ils ne minimisent pas la moyenne du carré des erreurs. La dérivation de modèles pour lesquels une décomposition optimale permet d'approcher un filtre d'ajustement de X11 ne pourra jamais reproduire cette propriété, et ne sera donc pas satisfaisante sous cet aspect (voir Burridge and Wallis (1984)). C'est une raison pour laquelle une approximation avec erreur nulle n'est pas réalisable. Un deuxième point important concerne la variabilité de court-terme des séries ajustées. Ce point est traité dans la section suivante.

5 Variabilité de court-terme des séries ajustées : filtres de X11 versus filtre de WK

Comparer des filtres ad hoc (X11) et des filtres reposant sur des modèles (WK) n'est jamais une tâche aisée. Une raison pour cela est que le filtre de WK requiert la spécification d'un modèle à partir duquel le filtre peut être dérivé. Les filtres ad hoc, par définition, n'exigent pas de tels préalables. Deux points doivent alors être soulignés. D'abord, si des arguments peuvent être *illustrés* avec ce modèle, les résultats ne doivent pas dépendre du choix du modèle. Ensuite, afin de rendre possible les comparaisons, les propriétés stochastiques représentées par ce modèle doivent être en accord avec les caractéristiques que les filtres de X11 sont supposés supprimer. En particulier, la structure de "band-pass" du filtre saisonnier doit être bien adaptée aux pics saisonniers du spectre associé au modèle considéré. Les filtres ad hoc sont ainsi plaçés dans une situation convenable, et la comparaison peut être conduite régulièrement. Un candidat satisfaisant ces points a été obtenu au travers

de la série PTIF (*voir figure 1 en annexe*), dont les propriétés sont bien décrites par le modèle suivant :

$$\Delta\Delta_{12}x_t = (1-.42B)\left(1-.61B^{12}\right)a_t . \qquad (5.1)$$

Aucun point aberrant n'a été trouvé et aucune transformation logarithmique n'a été nécessaire. Les résidus de ce modèle ne semblent pas corrélés : la statistique de Ljung-Box calculée sur 24 autocorrélations et la statistique de Box-Pierce sur les deux premiers retards saisonniers donnent respectivement un résultat de 24.66 et .68, non significatif dans les deux cas. L'indépendance semble aussi vérifiée puisque les mêmes statistiques calculées sur les carrés des résidus donnent respectivement un résultat de 20.47 and 5.24 respectivement, toujours non significatifs. Les statistiques de normalité confirment l'indépendance des résidus, avec une mesure de skewness égale à .25 (écart-type : .21) et une mesure de kurtosis égale à 2.97 (écart-type : 42). La variance des résidus s'élève à V_a = 1.847. Chaque coefficient du polynôme MA est très significativement différent de zéro.

Le modèle (5.1) appartient à la classe des modèles dits "airline", très couramment considérés dans la littérature depuis Box et Jenkins (1970). Les modèles "airline" ont l'avantage d'être simples et faciles à interpréter : les paramêtres des polynômes liés à la stabilité des composantes (voir par exemple Maravall, 1993). Avec les paramêtres fixés de cette façon, le modèle (5.1) a une propriété singulière : il est proche de la spécification que Burridge et Wallis (1984) ont proposée pour approcher le filtre d'ajustement par défaut de X11.

C'est le modèle (5.1) qui a été utilisé par les deux programmes X12-ARIMA et SEATS pour calculer les observations futures qui sont nécessaires à l'estimation des signaux près de la fin de l'échantillon. Etant donné que (5.1) ne demande pas de transformation logarithmique, la décomposition de la série PTIF est choisie comme additive dans les deux programmes. La décomposition a été menée en X12-ARIMA par le filtre d'ajustment par défaut : aucun problème particulier n'ayant été relevé dans les résultats. Dans l'approche AMB, l'ajustement saisonnier d'une série décrite par un modèle de type (5.1) est conduit de la façon suivante. Le spectre de la série observée est partagé selon les décompositions partielles de fractions en deux spectres, le premier capturant les mouvements aux fréquences saisonnières, le second les autres mouvements (voir Burman (1980)). Une décomposition unique est identifiée en enlevant autant de bruit blanc que possible de la composante saisonnière, selon le critère canonique (voir Hillmer et Tiao, 1982). Ce bruit blanc est affecté à la partie nonsaisonnière de la série. Pour le modèle (5.1), le spectre de la composante non saisonnière est associé au modèle :

$$\Delta^2 n_t = \left(1 - 1.39B + .41B^2\right)a_{nt}, \qquad (5.2)$$

avec $V_n = .67V_a$. Le filtre de WK est alors directement obtenu selon (4.2).

Le gain des filtres saisonniers de X11 et de WK sont présentés ensemble sur la *figure 3* (en annexe). Pour ce modèle (5.1), les deux filtres montrent la même structure de gain : ils laissent inchangées les mêmes gammes de fréquences et suppriment les autres. Par conséquent, les spectres des estimateurs de la saisonnalité présentés en *figure 4* (en annexe) sont pratiquement identiques. Tous deux semblent en accord avec le spectre de la série observée.

La *figure 5* (en annexe) montre les gains des filtres d'ajustement saisonniers. Les différences sont ici bien plus remarquables. Elles apparaissent en particulier entre les fréquences saisonnières harmoniques, et sont en relation avec le fait que le filtre d'ajustement de X11 a un gain plus grand que 1 à certaines fréquences tandis que le gain du filtre de WK ne peut jamais dépasser 1. Les fréquences où le filtre de X11 a un gain plus grand que 1 correspondent aux mouvements de court-terme de la série. Cela implique que la série ajustée sera sujette à des variations de court-terme plus larges que la série elle-même : les mouvements de court-terme sont amplifiés.

Le *figure 6* (en annexe) montre les spectres des estimateurs de la composante nonsaisonnière et le spectre de la série observée. La décomposition est en accord avec les propriétés stochastiques de la série telles qu'elles sont décrites par le modèle (5.1). Les deux estimateurs ont des spectres très proches l'un de l'autre, avec des différences entre les fréquences saisonnières harmoniques. Dans ces régions, le spectre de la série ajustée par X11 peut dépasser le spectre de la série observée. La série ajustée avec X11 présente donc une plus grande variabilité de court-terme que la série elle-même !

Puisque les variations de court-terme sont habituellement associée avec la composante irrègulière, il est naturel d'inspecter les estimateurs de cette composante. Les spectres des estimateurs obtenus avec X11 et SEATS sont donnés en *figure 7* (en annexe). L'amplification des mouvements de court-terme par X11 est évidente après la deuxième fréquence saisonnière. Il peut être facilement verifié sur l'aire de ces spectres que l'irrégulier selon X11 a une variance théorique plus large ($.23V_a$) que l'irrégulier selon l'approche AMB ($.19V_a$). Ce résultat est obtenu bien que l'approche AMB utilise l'hypothèse canonique qui maximise la variance de la composante irrégulière contenue dans la série.

Afin de vérifier si ces traits sont perceptibles en pratique, la *figure 8* (en annexe) présente la composante irrégulière estimée avec les deux approches. Une plus grande variabilité dans l'irrégulier de X11 est clairement observée : l'estimateur obtenu avec X11 a une variance de $.49=.27V_a$ contre $.32=.17V_a$ pour l'estimateur AMB.

L'amplification dans la série ajustée des mouvements de court-terme de la série originale a quelques incidences sur les taux de croissance. La *figure 9* (en annexe) montre les taux de croissance mensuels (in %) de la série PTIF ajustée obtenue avec X11 et l'approche AMB. Ils sont très proches en ce qui concerne la direction des mouvements, mais les taux de croissance mensuels de X11 sont plus erratiques, avec un écart-type de 1.14 contre 1.08 pour SEATS (points de pourcentage). Bien entendu, cette différence liée aux mouvements de court-terme serait atténuée si les taux de croissance étaient calculés sur plus longue période, telle que l'année par exemple.

La simulation d'une série selon (5.1) avec un échantillon de 300 observations a permis de vérifier que les conclusions de cette analyse ne sont pas affectés par un effet de fin d'échantillon. Cependant, il est intéressant de noter que Wallis (1982) a donné le gain du filtre d'ajustment de défaut de X11 pour l'estimation contemporaine : la particularité de présenter des gains plus grand que 1 à certaines fréquences y est bien plus marquée. Les filtres 3x9 et 3 termes présentent aussi des gains plus grand que 1 dans la région des huates fréquences d'une façon plus marquée (*voir figure 2 en annexe*). Les résultats présentés ici pourraient donc être plus apparent encore avec les filtres 3x9 ou 3 termes. Cette caractéristique des filtres de X11 se retrouve encore dans les filtres trimestriels (voir par exemple Bell et Monsell, 1992). Seul le filtre 3x3 ne partage pas cette propriété.

Bibliographie

Bell, W.R. (1984), « Signal Extraction for Nonstationary Time Series », *The Annals of Statistics*, 12, 2, 646-664.

Bell, W.R. et Monsell, B.C. (1992), « X-11 Symmetric Linear Filters and their Transfer Functions », Bureau of the Census, Research Report n. RR 92/15, Washington.

Box, G.E.P et Jenkins, G.M (1970), *Time Series Analysis : Forecasting and Control*, San Francisco : Holden Day.

Burman, J.P. (1980), « Seasonal Adjustment by Signal Extraction », Journal of the Royal Statistical Society, Ser. A, 143, 321-337.

Burridge, P. et Wallis, K.F. (1984), « Unobserved Component Models for Seasonal Adjustment Filters », *Journal of Business and Economic Statistics*, 2, 350-359.

Cleveland, W.P. et Tiao, G.C. (1976) « Decomposition of Seasonal Time Series : a Model for the X-11 Program », *Journal of the American Statistical Association*, 71, 581-587.

Fiorentini, G. et Planas, C. (1996), « Non-Admissible Decompositions in Unobserved Components Models », *Working Paper 96/13*, Cemfi, Madrid.

Hillmer S.C. , et Tiao G.C. (1982) « An ARIMA-Model-Based Approach to Seasonal Adjustment », *Journal of the American Statistical Association*, 77, 63-70.

Hylleberg, S. (1986), *Seasonality in Regression*, New York : Academic Press.

Maravall, A. (1993a), « Stochastic Linear Trends : Models and Estimators », *Journal of Econometrics*, 54, 1-33.

Maravall, A. (1993b), « Unobserved Components in Economic Time Series », prepared for the Handbook of Applied Econometrics, Vol. 1.

Maravall, A. et G\'{omez, V. (1992), « Signal Extraction in Economic Time Series : Program SEATS », EUI Working Paper ECO No. 92/65, Department of Economics, European University Institute.

Maravall, A. et Pierce, D.A. (1983), « Preliminary-Data Error and Monetary Aggregates », *Journal of Business and Economic Statistics*, 1, 179-186.

Maravall, A. et Pierce, D.A. (1986), « The Transmission of Data Noise into Policy Noise in U.S. Monetary Control », *Econometrica*, 54, n. 4, 961-979.

Pierce, D.A. (1979), « Signal Extraction Error in Nonstationary Time Series », *Annals of Statistics*, 7, 1303-1320.

Wallis, K.F. (1974), « Seasonal Adjustment and Relations between Vairables », *Journal of the American Statistical Association*, 69, 18-31.

Wallis, K.F. (1982), « Seasonal Adjustment and Revisions of Current Data : Linear Filters for the X11 Method », *Journal of the Royal Statistical Society*, Ser. A, 145, 74-85.

Whittle, P. (1963), *Prediction and Regulation using Least-Squares Methods*, London : English Universities Press.

Young, A.H. (1968), « Linear Approximation to the Census and BLS Seasonal Adjustment Methods », *Journal of the Royal Statistical Society*, Ser. A, 145, 74-85.

Figure 1 Production totale de l'industrie française (hors construction) (1985-1 à 1996-4)

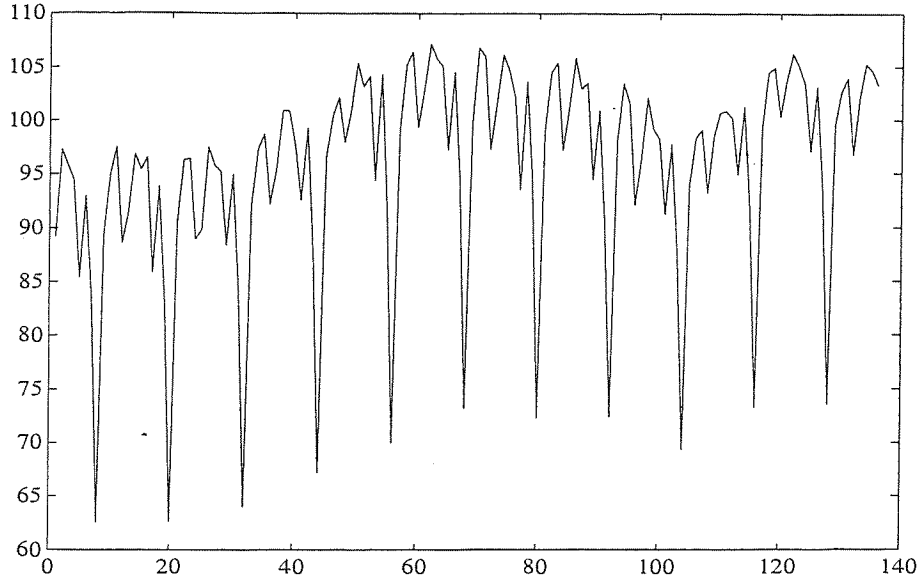


Figure 2 Fonctions de gain des filtres d'ajustement mensuel de X11

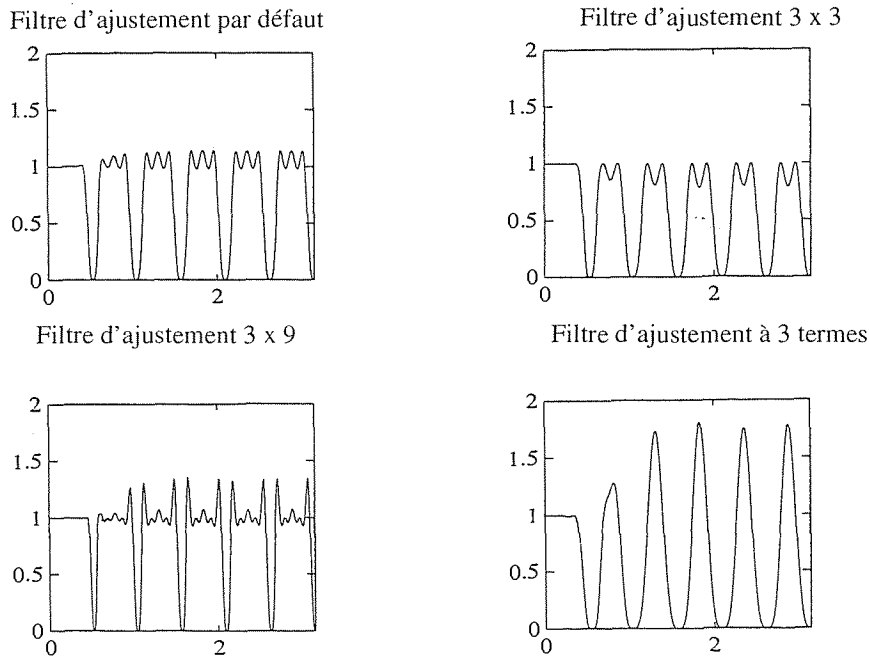


Figure 3 Gain des filtres saisonniers centraux : WK (—), X11 (- -)

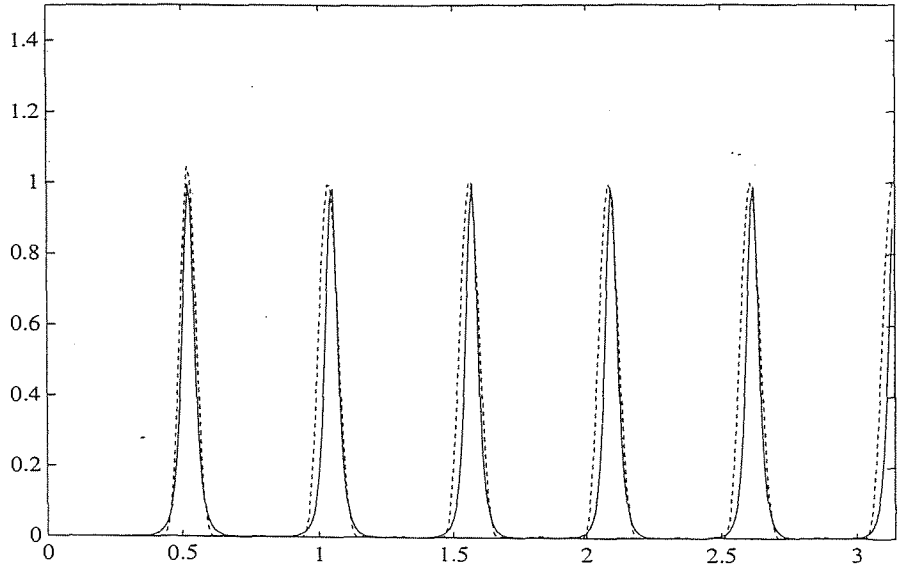


Figure 4 Spectres des estimateurs de la saisonnalité : WK (- -), X11 (- ·), spectre de la série (—)

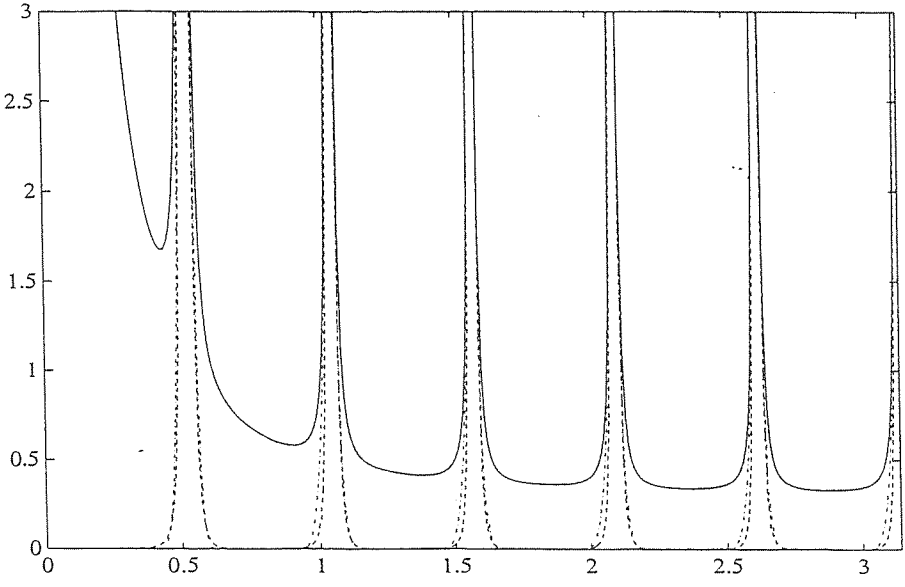


Figure 5 Gain des filtres centraux d'ajustement : WK (—), X11 (- -)

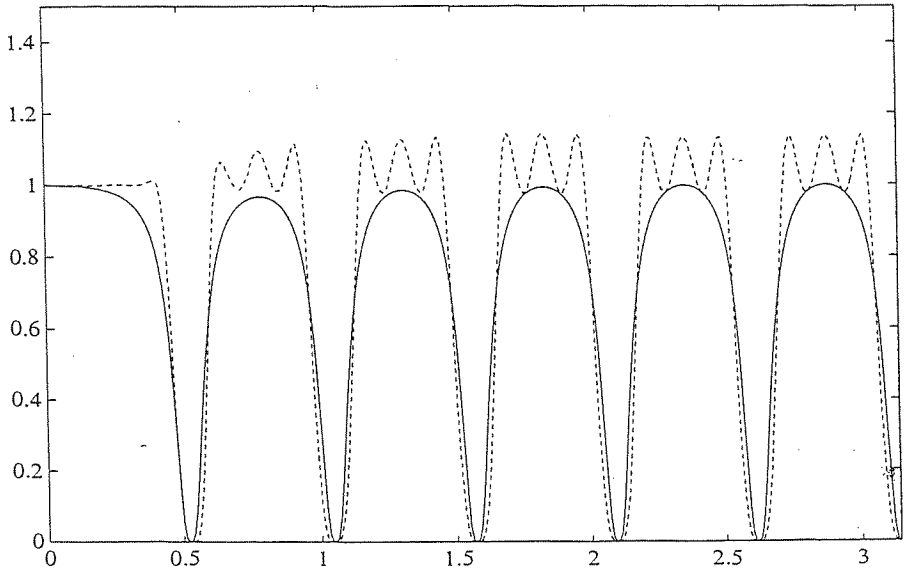


Figure 6 Spectres des estimateurs des séries ajustées : WK (- -), X11 (- ·), spectre de la série (—)

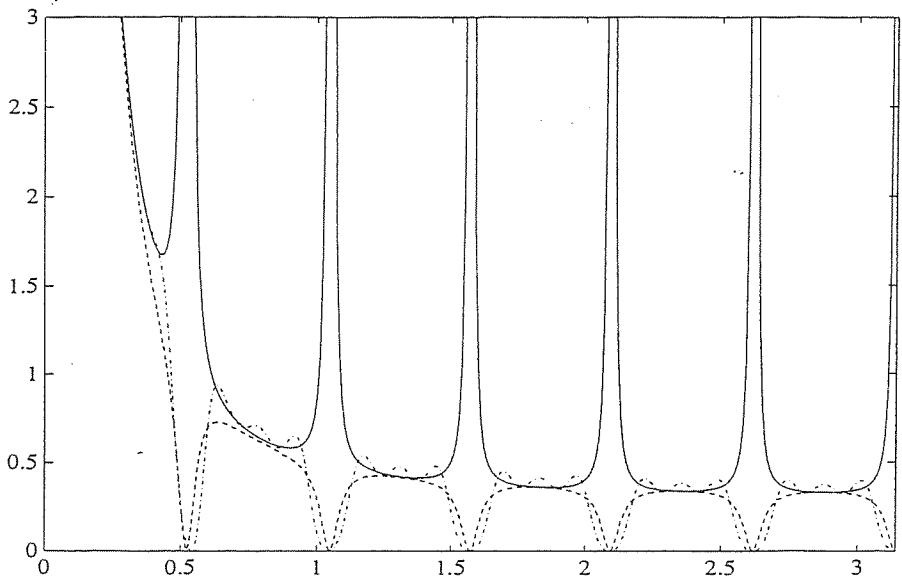


Figure 7 Spectres des estimateurs de la composante irrégulière : WK (—), X11 (- -)

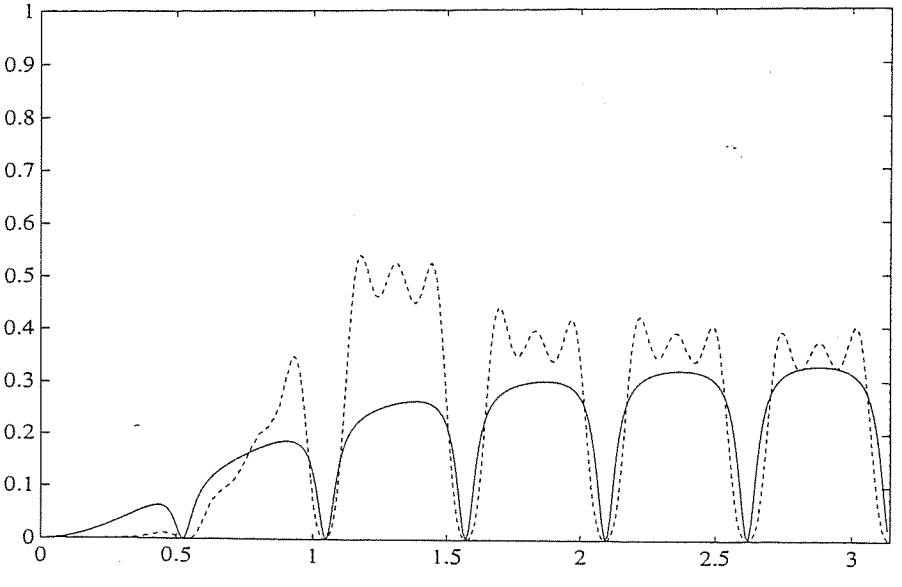


Figure 8 Estimateurs de la composante irrégulière

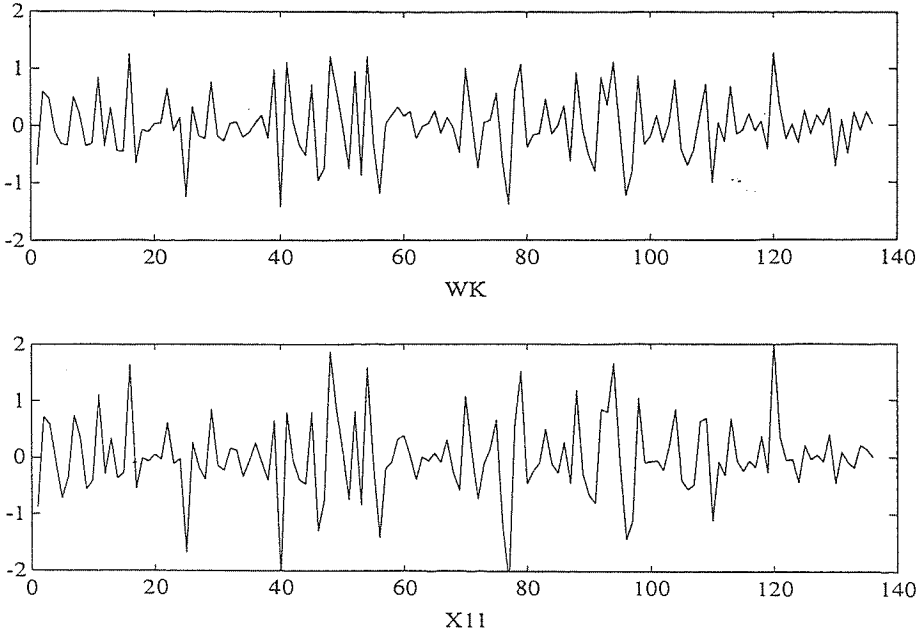
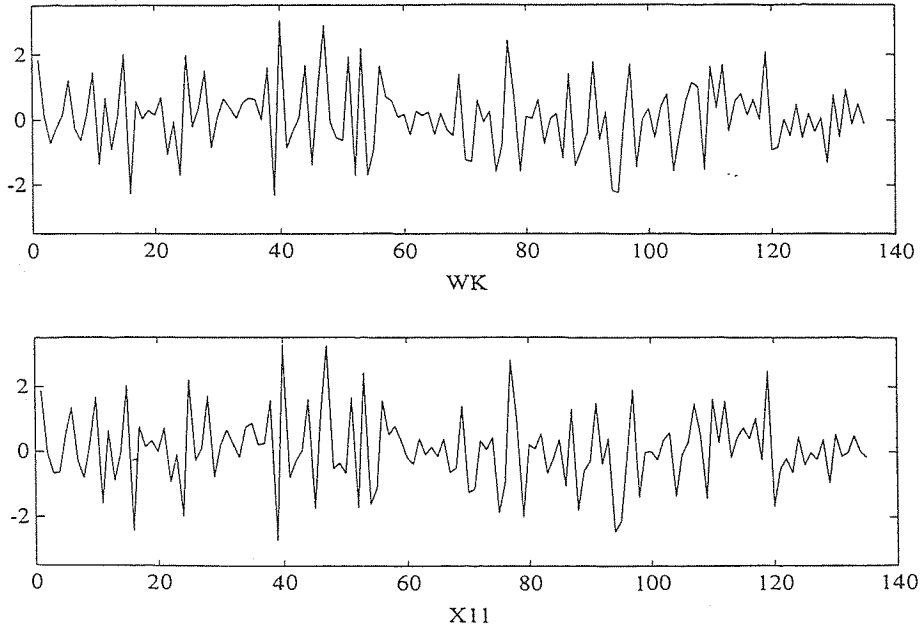


Figure 9 Taux de croissance mensuels de la série ajustée de la production (en %)



ANALYSE FACTORIELLE ET MODÈLES À COMPOSANTES INOBSERVABLES : APPLICATION À L'ÉTUDE DE L'ENQUÊTE DE CONJONCTURE DANS L'INDUSTRIE

Catherine Doz et Fabrice Lenglart

1 Introduction

Les enquêtes de conjoncture offrent un type d'information particulièrement utile pour l'analyse conjoncturelle. En effet, parmi l'ensemble des indicateurs infra-annuels, elles cumulent nombre d'avantages : elles fournissent un message recueilli directement auprès des acteurs économiques ; elles sont disponibles relativement rapidement (environ un mois après l'envoi des questionnaires) ; elles sont publiées à intervalles rapprochés et réguliers ; enfin, elles ne sont pas révisées.

Cependant, le nombre et la diversité des questions posées rendent souvent délicate l'interprétation des résultats obtenus. C'est pourquoi il semble utile de chercher à construire un indicateur unique à partir de l'ensemble des réponses fournies par une enquête, qui pourrait constituer une sorte de résumé de l'information qu'elle contient. Une façon simple de le faire consiste à calculer la moyenne simple de plusieurs soldes d'opinion relatifs à l'état de l'activité économique. C'est ainsi que l'institut allemand IFO publie un « indicateur des affaires », qui est la moyenne des soldes d'opinion concernant la production passée et futures des entreprises interrogées. Un autre exemple est celui de l'indice NAPM américain, résultat de la moyenne de cinq soldes concernant les entrées de commandes, la production, les stocks, l'emploi et les délais de livraisons. De telles méthodes présentent l'avantage de la simplicité, mais le choix des questions retenues comme celui des pondérations attribuées à chaque question demeurent *ad hoc*.

Une approche plus satisfaisante consiste à supposer que chaque variable peut être décomposée en deux composantes orthogonales entre elles : l'une commune à l'ensemble des séries, et l'autre spécifique à la variable considérée. La composante commune est alors assimilée à l'indice composite recherché. Les modèles à facteurs dynamiques (*dynamic factor models*) constituent le cadre d'analyse approprié pour formaliser de telles hypothèses. Ce type de modèles a été utilisé dans des contextes variés par Geweke (1977), Sargent et Sims (1977), Geweke et Singleton (1981),

Engle et Watson (1981, 1983), Watson et Kraft (1984) ou, plus récemment, par Stock et Watson (1989, 1991, 1993), Quah et Sargent (1993), Forni et Lippi (1995).

Deux méthodes peuvent être utilisées pour estimer ce genre de modèles. La première se place dans le domaine des fréquences. Dans ce cas, la dynamique du modèle n'a pas besoin d'être spécifiée : les méthodes standard d'analyse factorielle (encore appelés analyse en composantes communes et spécifiques (AFCS)) peuvent être utilisées, de façon à décomposer la matrice de densité spectrale. La deuxième méthode relève plus directement du domaine temporel : dès lors que la dynamique des différentes composantes a été spécifiée, le modèle peut être mis sous une forme espace-état et estimé par le filtre de Kalman.

Dans cette étude, nous utilisons des représentations ARMA pour modéliser les différentes composantes et nous mettons en oeuvre la deuxième méthode d'estimation. La représentation espace-état présente l'avantage d'être très « souple » : en particulier, elle permet d'utiliser simultanément des données à périodicités différentes, mensuelle et trimestrielle.

Néanmoins, nous appliquons également aux données la technique standard de l'analyse factorielle. Certes, cette méthode n'est pas *a priori* appropriée dans un cadre d'analyse dynamique (elle a été créée au départ pour étudier des données individuelles). Mais nous montrons qu'elle fournit des estimateurs convergents des paramètres du modèle, même dans le cas où il y a présence d'autocorrélation temporelle des variables et où cette autocorrélation n'est pas prise en compte. En définitive, les programmes d'analyse factorielle standard peuvent donc être utilisés. De surcroît, ils offrent des éléments statistiques qui aident à choisir le nombre de facteurs communs à de retenir. Ils offrent également des procédures de rotations d'axes qui facilitent l'interprétation des résultats obtenus lorsque plusieurs facteurs communs sont nécessaires pour décrire les données.

Les résultats obtenus par l'une ou l'autre méthode sont toujours très proches, ce qui renforce leur crédibilité. Nous commençons par estimer le modèle sur l'enquête mensuelle de conjoncture de l'INSEE dans l'industrie. Il apparaît qu'un seul facteur commun suffit à rendre compte de l'évolution commune des données, de sorte qu'il constitue un indicateur composite de l'enquête. Cet indicateur est estimé par les deux méthodes, et nous utilisons les deux estimations indifféremment pour commenter les fluctuations économiques dont elles rendent compte. Un tel indicateur peut être vu comme une sorte d'indicateur de « climat des affaires ». Dans une deuxième étape, nous tentons d'interpréter les mouvements spécifiques à chaque question de l'enquête, ce qui permet d'affiner le diagnostic conjoncturel. Enfin, nous commentons les résultats obtenus lorsque l'on mélange les enquêtes mensuelle et trimestrielle dans l'industrie.

La dernière partie du papier analyse plus à fond la structure d'information contenue dans l'indicateur synthétique, en étudiant les soldes d'opinion à un niveau plus désagrégé, celui des grandes branches industrielles (biens intermédiaires, biens d'équipement, biens de consommation). Des techniques de rotation axiale permettent de montrer que cet indicateur de climat n'est pas la résultante fortuite de conjonctures sectorielles indépendantes les unes des autres, mais qu'il est en fait présent dans chaque branche, même s'il s'y reflète suivant des modalités différentes.

2 Le cadre d'analyse théorique

2.1 Présentation de l'analyse factorielle standard

L'analyse factorielle a pour but d'offrir une description parcimonieuse d'un ensemble d'observations. Elle tente de représenter les variables étudiées dans un cadre linéaire, en fonction d'un ensemble de variables latentes, appelées *facteurs*, ou *facteurs communs*. C'est une technique qui est appropriée lorsqu'un petit nombre de facteurs peut rendre compte d'un part importante de l'information contenue dans l'ensemble des variables initiales.

Plus précisément, soit I le nombre de variables étudiées, soit T le nombre d'observations dont on dispose pour chaque variable, et soit y_{it} la valeur prise par la variable y_i à la date t [1]. Un modèle décrivant les variables $y_1, \ldots, y_I$ en fonction de J facteurs communs $F_1, \ldots F_J$, $J < I$, s'écrit :

$$y_{it} = \lambda_{i1} F_{1t} + \ldots + \lambda_{iJ} F_{Jt} + u_{it}, \text{ for } i = 1 \text{ to } I,$$

où les processus $\left(u_{it}\right)_{t \in Z}$ sont supposés indépendants entre eux et independant des facteurs, et où toutes les variables sont supposées centrées. Dans le cadre de l'analyse factorielle classique, *chacun des processus* $\left(F_{it}\right)$ *et* $\left(u_{it}\right)$ *est supposé sans autocorrélation* (nous abandonnerons cette restriction dans la section 2.2).

Ce modèle peut s'écrire de façon plus concise, en utilisant des notations matricielles.

---

[1] Comme nous l'avons précisé en introduction, ce type de modèle est généralement utilisé pour étudier I variables observées sur un échantillon de N individus, mais nous nous situons ici d'emblée dans un contexte temporel.

Soient : $y_t = (y_{1t}, \dots y_{It})'$, $F_t = (F_{1t}, \cdots F_{Jt})'$, $u_t = (u_{1t}, \dots, u_{It})'$, et $\Lambda = (\lambda_{ij})_{\substack{1 \le i \le I \\ 1 \le j \le J}}$.

On en déduit :

$$y_t = \Lambda F_t + u_t ,$$

avec : $EF_t = 0$, $Eu_t = 0$, $E(u_t u_t') = D = diag(\sigma_1^2, \dots \sigma_I^2)$,

$$\forall (t,\tau) \quad E(F_t u_\tau') = 0 ,$$

$$\forall (t,\tau), t \neq \tau, \quad E(F_t F_\tau') = 0$$

$$\forall (t,\tau), t \neq \tau, \quad E(u_t u_\tau') = 0$$

Un tel modèle n'est pertinent que lorsque les variables sont fortement corrélées entre elles ; dans ce cas, il permet d'analyser cette structure de corrélation. Plus précisément, les facteurs rendent compte des corrélations entre variables, alors que chaque u_{it} représente une source de variation affectant la seule variable y_{it} : chaque u_{it} est appelé *facteur* ou *composante spécifique* de la variable y_i. Les λ_{ij} sont appelés les *pondérations* (*loadings*) des facteurs communs : chaque λ_{ij} représente la contribution du facteur F_j à l'évolution de la variable y_i.

Il apparaît clairement que, dans une telle formulation, les facteurs ne sont définis qu'à une transformation linéaire près, sous réserve de modifier les pondérations. Dans le modèle le plus classique, les facteurs sont supposés non corrélés entre eux (cette hypothèse peut être ensuite levée, voir plus loin section 4.1) et avoir des variances unitaires (ce qui ne restreint pas la généralité du modèle puisqu'ils n'interviennent qu'à un facteur multiplicatif près). Cependant, même dans ce cas de figure, s'il y a plus d'un facteur à estimer, les facteurs ne sont pas définis de façon univoque : ils peuvent être modifiés via n'importe quelle rotation. Dans la pratique, lors de l'estimation, une solution unique est obtenue en ajoutant des contraintes identifiantes sur les facteurs. Cette solution peut ensuite subir une rotation (ou une simple transformation linéaire) si une telle opération facilite l'interprétation des résultats obtenus.

Concentrons nous maintenant sur le modèle de base, c'est-à-dire sur le cas où les facteurs sont non corrélés entre eux et sont de variance unitaire. Ce modèle conduit

à une interprétation simple en terme de variance et covariances des variables. De fait, on a dans ce cas :

$$y_t = \Lambda F_t + u_t ,$$

$$\text{avec } EF_t = 0 ,\ Eu_t = 0 ,\ E\left(F_T F_t\right) = Id ,\ E\left(u_t u'_t\right) = D ,$$

$$\forall (t,\tau)\ \ E\left(F_t u_\tau '\right) = 0 ,$$

$$\forall (t,\tau), t \neq \tau ,\ \ E(F_t\ F_\tau ') = 0$$

$$\forall (t,\tau), t \neq \tau ,\ \ E(u_t\ u_\tau ') = 0$$

On obtient alors les relations suivantes :

$Vy_t = \Lambda' \Lambda + D$,

et : $Vy_{it} = \sum_{j=1}^{J} \lambda_{ij}^2 + \sigma_i^2$ pour $i = 1$ à I, et pour tout t.

Chaque λ_{ij}^2 represente la part de la variance de y_{it} qui est expliquée par le facteur F_j, et $h_i^2 = \sum_{j=1}^{J} \lambda_{ij}^2$ represente la contribution totale des facteurs à la variance de y_{it} (h_i^2 est quelquefois appelée *communauté* (*communality*) de la variable y_i). La variance de u_{it}, σ_i^2, apparaît clairement comme la part de la variance de y_{it} qui ne peut pas être expliquée par les facteurs, puisque : $Vy_{it} = h_i^2 + \sigma_i^2$. Enfin, $V_j = \sum_{i=1}^{I} \lambda_{ij}^2$ mesure la contribution totale du facteur Fj aux variances de l'ensemble des variables.

Il existe deux méthodes principales pour estimer le modèle : l'analyse factorielle principale (*principal factor analysis*) et la méthode du maximum de vraisemblance sous hypothèse de normalité (voir Lawley et Maxwell (1971) pour une présentation complète). La première ne nécessite aucune connaissance sur le nombre de facteurs à retenir, alors que ce nombre doit être donné *a priori* pour mettre en oeuvre la seconde. En revanche, la méthode de maximum de vraisemblance fournit des estimateurs convergents des paramètres, ce qui n'est pas le cas de l'analyse factorielle principale.

En pratique, les deux méthodes sont utilisées. Lors d'une première étape, l'analyse factorielle principale fournit un critère permettant de choisir le nombre de facteurs à retenir. En fait, sa mise en oeuvre revient à effectuer une analyse en composantes principales (ACP) sur une matrice particulière : les termes hors diagonale de cette matrice sont égaux aux corrélations entre variables, mais les éléments diagonaux sont fixés, en première approximation, aux valeurs des corrélations canoniques de chaque variable avec l'ensemble des autres. On est alors conduit à calculer les valeurs propres de cette matrice, et le nombre de facteurs à retenir est choisi en fonction de la taille de ces valeurs propres. Dans une deuxième étape, il est alors possible de mettre en oeuvre la méthode de maximum de vraisemblance. En outre, un test du rapport de vraisemblance permet de contrôler que le nombre de facteur retenu est correct.[2]

Signalons enfin que, pour ce qui est des deux méthodes, les paramètres λ_{ij} sont estimés en premier, et que les valeurs prises par les facteurs communs F_{jt} (les *scores*) sont, dans un deuxième temps, approximés comme des combinaisons linéaires des variables initiales - ceci est fait par le biais de techniques de régression visant à minimiser la variance de l'écart entre chaque facteur et son approximation linéaire.

Dans ce papier, nous nous intéresserons particulièrement au cas où un facteur suffit pour avoir une bonne approximation de l'ensemble des variables observées. En effet, dans ce cas, cela a un sens de considérer le facteur F_t comme une sorte d'indicateur coïncident, puisqu'il rend compte d'une part importante de la variance de chaque variable. Le modèle prend alors la forme plus simple suivante :

$y_{it} = \lambda_i \, F_t + u_{it}$, pour tout i et tout t,

c'est-à-dire :

$y_t = \lambda \, F_t + u_t$,

où : $y_t = (y_{1t}, \ldots y_{Nt})'$, $\lambda = (\lambda_1, \ldots \lambda_N)'$, $u_t = (u_{1t}, \ldots u_{Nt})'$

$EF_t = 0$, $Eu_t = 0$, $VF_t = I$, $E(u_t \, u_t') = D$,

$\forall (t, \tau) \; E(F_t u_\tau) = 0$,

---

2 Ce test n'est cependant plus valide dans le cas où les processus sont temporellement autocorrélés.

$\forall (t,\tau),\ t \neq \tau,\ E(F_t F_\tau) = 0$

$\forall (t,\tau),\ t \neq \tau,\ E(u_t u_\tau{}') = 0$

2.2 Convergence, dans un cadre dynamique, des estimateurs obtenus dans l'analyse factorielle classique (statique)

On suppose que chacun des processus réels (F_{it}) et (u_{it}) peut présenter de l'autocorrélation, mais que le modèle est estimé par une procédure standard de maximum de vraisemblance, comme s'il n'y avait pas d'autocorrélation.

Soit θ le vecteur des paramètres : $\theta = (\lambda_{ij}, i = 1...I, j = 1,...J, \sigma_i^2, i = 1...I)$.

L'estimateur $\hat{\theta}_T$ ainsi obtenu est alors un *M*-estimateur de θ. Nous montrons que cet estimateur est convergent (les démonstrations des différents lemmes sont renvoyés en annexe). Il en résulte que les scores calculés à l'aide de cet estimateur ont les mêmes propriétés asymptotiques d'optimalité que dans le cas classique sans autocorrélation. En revanche, dans un tel contexte, on ne peut choisir le nombre de facteurs à retenir à l'aide de la procédure de test standard - le calcul de la matrice de variance-covariance asymptotique de $\hat{\theta}_T$, peut permettre d'élaborer un test adapté dans ce contexte, mais nous laissons ce point pour une recherche ultérieure.

Soit A la matrice de variance-covariance empirique des observations, et soit $C = \Lambda\Lambda' + D$ la matrice de variance-covariance théorique. La vraisemblance du modèle, calculée sous hypothèse de normalité, lorsque les facteurs et les composantes spécifiques ne présentent pas d'autocorrélation, s'écrit (à un terme constant près) : $\sum_t Ln\, l_t = -\frac{T-1}{2} Ln\, det\, C - \frac{T-1}{2} tr\, C^{-1} A$

(voir par exemple Lawley et Maxwell (1971)).

Soit θ_0 la vraie valeur du paramètre θ. On suppose que θ_0 vérifie :

$\theta_0 = (\lambda_{ij0}, i = 1...I, j = 1,...J, \sigma_{i0}^2, i = 1...I)$, avec $\sigma_{i0}^2 \neq 0$, pour $i = 1...I$.

Nous montrons ci-dessous que, si θ_0 vérifie cette hypothèse, alors le *M*-estimateur $\hat{\theta}_T$, obtenu par maximisation de la pseudo-vraisemblance précédente, vérifie un ensemble de conditions qui suffit à assurer sa convergence.

Dans toute la suite, on suppose que θ *varie dans une région de la forme* $R^I \times [\alpha, +\infty[^I$ *contenant* θ_0. Ceci entraîne notamment que θ vérifie lui-même : $\sigma_i^2 \neq 0$, pour $i = 1 \ldots I$. Cette dernière condition est en particulier une condition suffisante pour que C soit inversible, donc pour que la pseudo-vraisemblance soit définie.

Pour tout t fixé, on note $x_{it} = y_{it} - \bar{y}_i$ et $x_t = (x_{1t}, \ldots x_{It})'$. La fonction à maximiser, que nous noterons par la suite $Q_T(y, \theta)$, s'écrit alors :

$$Q_T(y,\theta) = -\frac{1}{2} Ln\ det\ C - \frac{1}{2} tr\left(C^{-1} \left(\frac{1}{T} \sum_t x_t\ x_t' \right) \right)$$

Enfin, pour toute matrice M on note : $\|M\| = \underset{i,j}{Max} |m_{ij}|$.

Lemme 1 :

i) Si $\|C\| \to +\infty$, alors $Q_T(y, \theta) \to -\infty$.

Il résulte de ce lemme que, lorsqu'on cherche à maximiser $Q_T(y, \theta)$ sur une région de la forme $R^I \times [\alpha, +\infty[^I$, on peut se restreindre au cas où θ varie dans un compact inclus dans cette région.

Dans toute la suite, on suppose donc que θ appartient à un compact Θ contenant θ_0 et inclus dans une région de la forme $R^I \times [\alpha, +\infty[^I$, où α est un réel fixé.

Lemme 2 : $Q_T(y, \theta)$ converge en probabilité, uniformément pour $\theta \in \Theta$, vers :

$$Q_0(\theta) = -\frac{1}{2} Ln\ det\ C - \frac{1}{2} tr\ C^{-1}\ C_0,$$

où $C_0 = \Lambda_0 \Lambda_0' + D_0$ est la matrice de variance covariance associée à la vraie valeur θ_0 du paramètre.

Lemme 3 : La fonction Q_0 admet un maximum unique au point θ_0.

Lemme 4 : La fonction Q_0 est continue sur Θ.

Proposition : Le *M*-estimateur $\hat{\theta}_T$ est convergent.

Preuve : Par application du lemme 1, on a vu qu'on pouvait supposer que $\theta \in \Theta$, où Θ est un compact inclus dans $R^I \times \left(R^{+*}\right)^I$. Le résultat découle alors des lemmes 2 à 4 et des théorèmes généraux sur la convergence des *M*-estimateurs (voir par exemple Newey et Mac Fadden (1994), théorème 2.1).

2.3 Les modèles à composantes inobservables

L'analyse factorielle classique est une méthode rapide et très simple d'emploi. Cependant, elle présente l'inconvénient de ne pas préciser la dynamique des variables étudiées -même si, comme nous l'avons vu, cette méthode peut être utilisée dans un cadre dynamique - de sorte que, par exemple, elle ne peut pas être utilisée dans un but de prévision. De plus, les procédures standard d'analyse factorielle ne peuvent pas traiter le cas de variables dont les périodicités sont différentes.

Comme nous l'avons dit en introduction, dès lors que la dynamique des différentes composantes est spécifiée, un modèle dynamique à facteurs peut être écrit à l'aide d'une representation espace-état. Comme les facteurs sont des variables inobservables, un tel modèle appartient à la classe des modèles à composantes inobservables (M.C.I.), et peut être estimé en utilisant la technique du filtre de Kalman.

Plaçons nous dans la situation où un seul facteur commun suffit à expliquer une grande part de la variance des variables - ceci doit être le cas lorsqu'on veut construire un indicateur coïncident. Des travaux antérieurs nous ayant conduit à spécifier un modèle ARMA(2,1) pour le facteur commun, et des modèles AR(1) pour les composantes spécifiques, on obtient alors le modèle suivant :

$$y_{it} = \lambda_i F_t + u_{it}, \text{ pour } i = 1 \text{ à } I, \text{ pour tout } t,$$

$$F_t = \varphi_1 F_{t-1} + \varphi_2 F_{t-2} + \varepsilon_t - \theta \varepsilon_{t-1}, \text{ pour tout } t,$$

$$u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}, \text{ pour } i = 1 \text{ à } I, \text{ pour tout } t,$$

où ε_t et ε_{it} sont les innovations de F_t et de u_{it} à chaque date t, et où les processus (ε_t) et (ε_{it}) sont supposés indépendants.

Ce type de modèle admet une représentation espace-état. Rappelons que la forme générale des modèles espace-état est la suivante :

$$y_t = Z_t\,\alpha_t + d_t + e_t \qquad (M)$$
$$\alpha_t = A_t\,\alpha_{t-1} + c_t + R_t\,\eta_t \qquad (T)$$

dans laquelle les processus (e_t)et (η_t) sont non autocorrélés, non corrélés entre eux, et vérifient : $Ee_t = 0 \quad Ve_t = H_t \quad E\eta_t = 0 \quad V\eta_t = Q_t$.

Dans un tel cadre, le vecteur α_t, qui est appelé le *vecteur d'état*, est inobservable. L'équation *(M)* est appelée *équation de mesure* : elle donne la relation, à chaque date t, entre le vecteur y_t des variables observées et le vecteur d'état. L'équation *(T)*, ou *équation de transition*, décrit la dynamique du vecteur d'état.

Les matrices Z_t, A_t, d_t, c_t, R_t, H_t et Q_t sont généralement non stochastiques, mais elles peuvent dépendre du temps de façon déterministe. Elles peuvent aussi dépendre des paramètres du modèle.

Si l'on définit le vecteur d'état par : $\alpha_t = (F_t, F_{t-1}, \varepsilon_t, u_{1t}, \ldots u_{It})'$, on peut écrire le modèle sous forme espace-état en utilisant les relations suivantes :

$$y_t = \begin{pmatrix} \lambda_1 & 0 & 0 & 1 & & \\ \vdots & \vdots & \vdots & & \ddots & \\ \lambda_I & 0 & 0 & & & 1 \end{pmatrix} \begin{pmatrix} F_t \\ F_{t-1} \\ \varepsilon_t \\ u_{1t} \\ \vdots \\ u_{It} \end{pmatrix},$$

et :
$$\begin{pmatrix} F_t \\ F_{t-1} \\ \varepsilon_t \\ u_{1t} \\ \vdots \\ u_{It} \end{pmatrix} = \left(\begin{array}{ccc|ccc} \varphi_1 & \varphi_2 & -\theta & & & \\ 1 & 0 & 0 & & 0 & \\ 0 & 0 & 0 & & & \\ \hline & & & \rho_1 & & \\ & 0 & & & \ddots & \\ & & & & & \rho_I \end{array}\right) \begin{pmatrix} F_{t-1} \\ F_{t-2} \\ \varepsilon_{t-1} \\ u_{1,\,t-1} \\ \vdots \\ u_{I,\,t-1} \end{pmatrix} + \left(\begin{array}{c|ccc} 1 & & & \\ 0 & & 0 & \\ 1 & & & \\ \hline & 1 & & \\ 0 & & \ddots & \\ & & & 1 \end{array}\right) \begin{pmatrix} \varepsilon_t \\ \varepsilon_{1,\,t} \\ \vdots \\ \varepsilon_{It} \end{pmatrix}$$

Ceci peut s'écrire de façon plus concise, en utilisant les notations introduites précédemment, et en notant de plus : $\tilde{\varepsilon}_t = (\varepsilon_{1t}, \ldots \varepsilon_{It})'$, $R = diag\,(\rho_1, \cdots \rho_I)$, et $\underset{(I+3,\,1)}{\alpha_t} = (F_t, F_{t-1}, \varepsilon_t, u_t')'$. On obtient :

$$y_t = (\lambda \; 0 \; 0 \; Id) \begin{pmatrix} F_t \\ F_{t-1} \\ \varepsilon_t \\ u_t \end{pmatrix} \qquad (M)$$

$$\begin{pmatrix} F_t \\ F_{t-1} \\ \varepsilon_t \\ u_t \end{pmatrix} = \left(\begin{array}{ccc|c} \varphi_1 & \varphi_2 & -\theta & \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \\ \hline & 0 & & R \end{array}\right) \begin{pmatrix} F_{t-1} \\ F_{t-2} \\ \varepsilon_{t-1} \\ u_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & Id \end{pmatrix} \begin{pmatrix} \varepsilon_t \\ \tilde{\varepsilon}_t \end{pmatrix} \qquad (T),$$

où $(\varepsilon_t, \tilde{\varepsilon}_t')'$ a une matrice de variance-covariance diagonale.

Un tel modèle peut être estimé de diverses façons. Dans ce papier, nous utilisons une procédure du maximum de vraisemblance à l'aide de la technique du filtre de Kalman. Rappelons brièvement les aspects principaux de cette méthode - une présentation complète figure, par exemple, dans Harvey (1989) ou Gourieroux & Monfort (1990). Pour tout ensemble fixé de valeurs des paramètres, le filtre de Kalman en tant que tel consiste à déterminer la meilleure prévision du vecteur d'état à chaque date, compte tenu de l'information disponible à cette date. Pour tout t, les prévisions optimales : $a_{t|t-1} = E(\alpha_t | I_{t-1})$ et $a_t = E(\alpha_t | I_t)$, sont ainsi calculées par une procédure itérative.

Sous l'hypothèse de normalité des perturbations, il est alors possible de calculer la vraisemblance du modèle : $\ell(y_1, \ldots y_T, \psi) = \prod_{t=1}^{T} f(y_t / \mathrm{I}_{t-1})$, où $f(y_t / \mathrm{I}_{t-1})$ désigne la densité conditionnelle de y_t sachant $(y_1 \ldots y_{t-1})$, et où ψ désigne le vecteur des paramètres. En effet, $f(y_t / \mathrm{I}_{t-1})$ est alors une loi $N(Z_t\, a_{t|t-1}, W_t)$ et la matrice de variance-covariance W_t est une matrice qui est calculée au cours de la procédure du filtre de Kalman. La vraisemblance peut ainsi être obtenue en tout point de l'espace des paramètres, de sorte qu'on peut la maximiser à l'aide d'une procédure quelconque d'optimisation, de façon à obtenir les estimateurs du maximum de vraisemblance des paramètres. Sous des hypothèses standard, ces estimateurs sont asymptotiquement normaux.

Enfin, on peut réappliquer le filtre de Kalman en utilisant les valeurs estimées des paramètres, et on obtient ainsi l'approximation correspondante du vecteur d'état à chaque date t ; mais on peut aussi améliorer cette approximation et déterminer la prévision optimale du vecteur d'état à chaque date t, sachant l'ensemble de l'information disponible. Ceci consiste à calculer $E(\alpha_t | I_T)$, et ce calcul peut être mené à l'aide d'une nouvelle procédure itérative, appelée procédure de lissage de Kalman.

2.4 Application à des données de périodicités différentes

Les modèles à composantes inobservables offrent aussi un cadre bien adapté pour estimer un modèle à facteurs lorsque les données n'ont pas toutes la même périodicité. Considérons le cas où l'on s'intéresse à la fois à un ensemble de données mensuelles : $y_t = (y_{1t}, \ldots y_{I_1 t})'$, et à un ensemble de données trimestrielles : $z_t = (z_{1t}, \ldots z_{I_2 t})'$. On suppose que t désigne l'indice du mois et que z_t n'est observé que lorsque $t \equiv 0\,[3]$.

Nous faisons ici l'hypothèse que le même facteur commun peut rendre compte des deux ensembles de données. Avec les notations précédentes, le modèle à estimer est donc le suivant :

$$\underset{(I_1, 1)}{y_t} = \lambda\, F_t + u_t$$

$$\underset{(I_2, 1)}{z_t} = \mu\, F_t + v_t$$

où $\mu = (\mu_1, \ldots \mu_{I_2})'$, et $v_t = (v_{1t}, \ldots v_{I_2 t})'$.

Les procédures standard d'analyse factorielle ne permettent pas d'estimer un tel modèle. Cependant, si l'on spécifie la dynamique des variables, il est possible d'écrire un modèle à composantes inobservables, comme nous le montrons ci-dessous.

On spécifie ici la dynamique des variables sous la même forme que précédemment : F_t est un processus ARMA(2,1), les u_{it} et les v_{it} sont des processus AR(1). On conserve les mêmes notations que précédemment pour les dynamiques de F_t et des u_{it}, et on suppose que :

$v_{it} = \tau_i\, v_{i,t-1} + \zeta_{it}$, pour tout i, pour tout t,

où ζ_{it} est l'innovation de v_{it} et est non corrélé avec les autres perturbations.

On pose de plus : $\zeta_t = \left(\zeta_{1t}, \ldots \zeta_{I_2 t}\right)'$ et $T = \text{diag}\,(\tau_1, \ldots \tau_{I_2})$

Définissons maintenant une variable Y_t dont la taille varie selon la date :

$$Y_t = \begin{pmatrix} y_t \\ z_t \end{pmatrix} \text{ si } t \equiv 0\,[3],$$

$Y_t = y_t$ sinon.

On peut alors obtenir une représentation espace-état du modèle sous la forme suivante :

- $\alpha_t = (F_t, F_{t-1}, \varepsilon_t, u_t', v_t')'$,

- $Y_t = Z_t\,\alpha_t$ où : $Z_t = \begin{pmatrix} \lambda & 0 & 0 & \text{Id} & 0 \\ \mu & 0 & 0 & 0 & \text{Id} \end{pmatrix}$ si $t \equiv 0\,[3]$

$$Z_t = \begin{pmatrix} \lambda & 0 & 0 & \text{Id} & 0 \end{pmatrix} \text{ sinon,}$$

- $\alpha_t = A_t\,\alpha_{t-1} + R_t\,\eta_t$, où :

$$\eta_t = \begin{pmatrix} \varepsilon_t \\ \tilde{\varepsilon}_t \\ \zeta_t \end{pmatrix}, A_t = \left(\begin{array}{ccc|cc} \varphi_1 & \varphi_2 & -\theta & & \\ 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & & \\ \hline & 0 & & R & 0 \\ & & & 0 & T \end{array}\right), R_t = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & \text{Id} & 0 \\ 0 & 0 & \text{Id} \end{pmatrix},$$

et où η_t a une matrice de variance-covariance diagonale et est non-autocorrélé.

3 Application aux données de l'enquête de conjoncture dans l'industrie

3.1 Les données

On considère ici les réponses à six questions de l'enquête de conjoncture de l'INSEE dans l'industrie, sur la période janvier 1963-janvier 1996. Ces questions sont : opinion sur la tendance passée de la production personnelle (TPPA), opinion sur la tendance future de la production personnelle (TPPRE), opinion sur la demande et les carnets de commande globaux (OSCD), opinion sur la demande et les carnets de commande en provenance de l'étranger (OSCDE), opinion sur le niveau des stocks (OSSK) et perpectives générales d'activité (PGP).

Dans cette enquête, pour chaque question, les industriels peuvent répondre suivant trois modalités : le niveau est plus élevé, moins élevé, ou comparable à celui de la période précédente. Les réponses sont agrégées en pondérant par la taille de l'entreprise répondante. Le résultat publié est le solde d'opinion, c'est-à-dire la différence entre le pourcentage de réponses « plus élevé » et celui de réponses « moins élevé ».

Un simple coup d'oeil au *graphique 1* des six soldes (en annexe) suffit pour se convaincre que les profils de ces séries sont extrêmement similaires. D'où l'idée de chercher à résumer en un seul indice synthétique l'information commune qu'elles contiennent.

Chaque solde d'opinion présente une dynamique « cyclique » très prononcée, en ce sens qu'il peut être considéré comme stationnaire, mais qu'il présente une structure d'autocorrélation assurant une persistance à ses variations. En effet, les autocorrélogrammes empiriques montrent que les autocorrélations sont significativement positives jusqu'à l'ordre 8. Néanmoins, ces séries n'en sont pas pour autant intégrées : le test Dickey-Fuller augmenté refuse l'hypothèse nulle de la présence d'une racine unité dans la modélisation ARMA. Tout ceci justifie l'utilisation du modèle à composantes inobservables présenté plus haut, dans lequel toutes les variables sont stationnaires.

3.2 - Résultats généraux obtenus par les deux méthodes

L'analyse factorielle appliquée aux six soldes d'opinion conduit à retenir un seul facteur commun. En effet, ce facteur suffit à expliquer 82% de la variance totale des séries. Plus précisément, les pondérations des facteurs et les communautés sont :

Question	Pondérations	Communautés
TPPA	0.96	91.7 %
TPPRE	0.89	78.5 %
OSCD	0.97	93.2 %
OSCDE	0.89	79.2 %
OSSK	-0.88	77.1 %
PGP	0.86	73.6 %

Lorsqu'on estime le modèle dynamique à facteur présenté précédemment, en utilisant la technique du filtre de Kalman, on obtient les résultats suivants :

- les estimations obtenues pour les paramètres du facteur commun conduisent à l'équation suivante :

$$F_t = \underset{(0.04)}{1.88}\, F_{t-1} - \underset{(0.04)}{0.89}\, F_{t-2} + u_t - \underset{(0.07)}{0.60}\, u_{t-1}$$

(il faut ici noter que la valeur de σ_u a été fixée : $\sigma_u = 0.5$; en effet le modèle est défini à un paramètre d'échelle près, et il faut donc fixer la valeur de l'un des paramètres pour le rendre identifiable).

- les estimations des autres paramètres figurent dans le tableau suivant :

Question	λ_i		ρ_i		σ_i	
	estimation	écart-type	estimation	écart-type	estimation	écart-type
TPPA	3.10	0.18	0.69	0.03	3.35	0.09
TPPRE	2.34	0.21	0.86	0.03	3.07	0.10
OSCD.	4.00	0.20	0.88	0.04	2.57	0.14
OSCDE	3.50	0.26	0.92	0.02	4.00	0.16
OSSK	-2.37	0.16	0.82	0.04	3.18	0.11
PGP	5.08	0.38	0.89	0.03	6.72	0.19

On peut remarquer que l'estimation obtenue pour le polynôme autorégressif du facteur commun a des racines proches de l'unité. Cependant, il semble difficile de

tester directement la présence d'une racine unité dans un tel contexte. En effet, sous l'hypothèse nulle, la loi est non standard et les tests de Dickey-Fuller ne peuvent être directement appliqués dans le cadre non-linéaire des modèles à composantes inobservables. Cependant, les tests de racines unité menées sur les séries initiales ont conduit à rejeter l'hypothèse de non-stationnarité, et le facteur commun est ici nécessairement de même nature. Ce type de résultat n'est d'ailleurs pas spécifique à notre étude ; au contraire, il semble que l'on obtienne souvent des résultats analogues lorsqu'on étudie des séries présentant un fort caractère cyclique, ceci étant dû au caractère fortement persistant de telles séries[3].

Comme nous travaillons dans un cadre stationnaire, les statistiques usuelles peuvent être utilisées. Cependant, le problème de la significativité des paramètres d'écart-type est un peu plus complexe puisque la valeur zéro est sur la frontière de l'espace admissible des paramètres. Dans ce cas, le carré de la statistique de Student suit un mélange de lois : $1/2\chi^2_0 + 1/2\chi^2_1$[4]. L'étude des résultats conduit, ici encore, à conclure que ces paramètres sont tous très significatifs.

Notons enfin que les facteurs communs obtenus avec les deux techniques sont extrêmement proches (cf. *graphique 2* en annexe). Ceci nous conduit à les utiliser indifféremment pour commenter les fluctuations économiques qu'ils reflètent.

3.3 Interprétation du facteur commun

Lorsque l'on rapproche l'évolution de l'indicateur obtenu du glissement annuel de la production manufacturière (tirée des comptes trimestriels), on constate que les retournements des deux séries se produisent le plus souvent à des dates très voisines[5](cf. *graphique 3* en annexe). Il ne faudrait pas, cependant, se méprendre sur le sens de cette comparaison : on se borne à constater que le regard porté par les industriels sur l'activité de leur secteur (au travers de leurs réponses à l'enquête mensuelle) offre un reflet fidèle des grands mouvements et des inflexions de la

---

3 Par exemple, Watson (1986) estime la composante cyclique du PIB américain sous une forme AR(2) dans le cadre d'un M.C.I. et il obtient : $C_t = 1.501\ C_{t-1} - 0.577\ C_{t-2} + u_t$. Ici aussi, $\varphi_1 + \varphi_2 \approx 1$, ce qui signifie que les racine du polynome retard sont en module proches de l'unité. Watson relève ce fait, mais ne teste pas l'hypothèse de non stationnarité.

4 Voir par exemple Harvey (1989) : un test de niveau 5% est obtenu en comparant le Student à 1.6.

5 La comparaison, a priori plus naturelle, entre la variable synthétique et le glissement annuel de la production de l'ensemble de l'industrie fournit le même type d'enseignement, mais l'aspect heurté de l'activité de la branche énergie rend le rapprochement un peu moins lisible. Ceci explique qu'on ait choisi comme illustration le glissement annuel de la seule production manufacturière.

conjoncture industrielle. En outre, les variations au mois le mois de cette variable apparaissent moins heurtées que celles des soldes d'opinion relatifs aux différentes questions de l'enquête, ce qui permet en principe de détecter plus facilement et plus rapidement les inflexions qu'elle peut connaître.

A l'observer sur une période relativement longue, cet indicateur synthétique permet ainsi de suivre les principales fluctuations qu'a connues le secteur industriel français. On y lit, dans les années 60, 70 et au début des années 80, des variations souvent en partie dues aux alternances des politiques gouvernementales de relance et de freinage de l'activité, ainsi que l'impact des deux chocs pétroliers (en 1974-75 puis 1979-80). L'expansion de la fin des années 80, initiée par le contre-choc pétrolier, apparaît très forte, puis le climat général rend compte du ralentissement entamé dans le courant de 1990, marqué par le faux redémarrage du printemps 1992 et par la récession profonde de 1993. Enfin, la remontée en flèche de 1994 traduit la vigueur et la brièveté de la dernière reprise industrielle.

3.4 Interprétation des composantes spécifiques

L'indice composite ici calculé rend compte de la plus grande part de l'information contenue dans les différents soldes. Cependant, si l'on veut analyser toute l'information fournie par l'enquête, il faut également interpréter les composantes spécifiques de chaque question. En fait, l'intérêt des enquêtes de conjoncture est double : d'une part, les résultats fournis peuvent être utilisés pour tenter de construire des prévisions quantitatives concernant l'évolution de certains agrégats économiques ; d'autre part, ils contribuent à affiner le diagnostic conjoncturel. L'utilisation de nos travaux à des fins d'estimation quantitatives fera l'objet de travaux ultérieurs. Nous nous contentons ici de montrer que, durant les périodes où la composante spécifique d'un solde d'opinion présente un mouvement prononcé et durable, il est souvent possible d'interpréter cette information d'un point de vue qualitatif.

Il est très important ici de ne pas se laisser abuser par les expressions employées : le terme d'information spécifique désigne l'information complémentaire apportée par une question donnée ; en particulier, il ne signifie pas qu'il faille se fonder uniquement sur elle pour analyser les résultats fournis par cette question. Bien au contraire, on a vu que l'indicateur de climat, cette information commune à toutes les questions, est responsable d'une part très importante de l'évolution du profil de chacune d'entre elles. Si bien que, pour prendre un exemple, en période de climat favorable, si une composante spécifique apparaît négative de façon durable, on jugera, non pas que l'opinion des industriels sur le sujet abordé est défavorable, mais qu'elle est plutôt moins favorable qu'elle ne pourrait -ou ne devrait- l'être dans le contexte conjoncturel étudié. Cette précaution prise, on peut maintenant tenter, à la

lumière des réponses aux questions de l'enquête mensuelle, de préciser la description des fluctuations conjoncturelles que l'industrie française a traversées.

La réponse relative au **niveau de la production passée** (cf *graphique 4.1* en annexe)est très proche du climat général, si bien qu'elle constitue à elle seule une bonne approximation de l'information commune sous-jacente aux six questions de l'enquête, à condition de la lisser quelque peu. Cet avantage a pour contrepartie évidente le fait que l'information spécifique apportée par l'opinion sur l'activité passée, animée de mouvements de faible amplitude, généralement irréguliers et contradictoires, ne semble pas interprétable.

En ce qui concerne les **perspectives personnelles de production** (cf *graphique 4.2* en annexe), si l'on veut voir dans l'information spécifique fournie par cette question une variable anticipée de la conjoncture, le bilan que l'on peut dresser est pour le moins contrasté. En fait, il semble, en première approximation, que, jusqu'à la fin des années 70, cette information spécifique tendait à positionner le solde d'opinion au dessus du climat général en période de conjoncture ascendante et en dessous dans le cas contraire, si bien que le croisement des deux courbes signalait qu'un véritable retournement était en cours ; en ce sens, on pouvait donc estimer que les grandes inflexions de l'activité pouvaient être décelées au travers des anticipations des entrepreneurs. Ceci peut signifier que la moindre internationalisation d'alors de l'économie française et la plus grande efficacité des politiques gouvernementales de régulation rendaient plus aisée la compréhension de la conjoncture et de ses mouvements à venir. La situation semble s'être obscurcie depuis le début des années 80. Néanmoins, on peut noter que la composante spécifique semble avoir retrouvé un certain pouvoir prédicteur lors du dernier cycle (93-95).

C'est généralement en période de haut ou de bas de cycle que l'information spécifique apportée par la réponse sur le **niveau de la demande et des carnets globaux** (cf *graphique 4.3* en annexe) se distingue du climat général des affaires. Une interprétation possible est la suivante : lorsque l'information spécifique est positive, cela signifie que l'état de la demande adressée aux industriels interrogés est jugé meilleur que ce que laisserait prévoir le contexte conjoncturel qu'ils décrivent : ils font donc sans doute preuve d'un pessimisme personnel relatif lorsqu'ils rendent compte du climat général. Ce cas de figure semble avoir prévalu lors des années de forte expansion de 1989 et 1990. Si ce point de vue est juste, il faut mettre en regard cette période et l'année 1994, au cours de laquelle la composante est, au contraire, faiblement négative ; ceci signifierait que les industriels ont perçu, lors de la dernière reprise, une réalité plutôt plus favorable que ce qu'indiquaient les carnets, se montrant cette fois plutôt optimistes.

L'information spécifique liée à la question sur **la demande et les carnets étrangers** (cf *graphique 4.4* en annexe) offre un éclairage extrêmement intéressant sur les décalages conjoncturels éventuels entre la France et l'étranger. On y voit par

exemple que la France n'a pas bénéficié pleinement de la reprise mondiale de 1984. On peut néanmoins noter que cette composante a tendance sur l'espace de trente ans à se réduire, de sorte que les évolutions du solde d'opinion semblent se rapprocher globalement de celles suivies par le climat général. Ceci peut sans doute s'interpréter comme un signe tangible de l'ouverture progressive de l'économie, et plus spécifiquement du secteur industriel, vers l'extérieur, si bien que les fluctuations conjoncturelles majeures suivies par le pays apparaissent de plus en plus reliées à celles que traversent ses principaux partenaires. Depuis 1991, l'activité industrielle de la France apparaît, à cet égard, plutôt en phase avec celle de l'Europe occidentale considérée dans son ensemble.

Un intérêt de la décomposition en informations commune et spécifique du solde d'opinion relatif au **niveau des stocks** (cf. *graphique 4.5* en annexe) pourrait être de proposer une lecture plus précise de la réponse à cette question qui est, en effet, toujours difficile à analyser. On demande aux industriels s'ils jugent le niveau de leurs stocks supérieur, inférieur ou conforme à la normale. La difficulté porte donc sur l'interprétation à donner à ce terme : la "normale" désigne-t-elle un niveau moyen de stocks, constant au cours du temps, ou bien un niveau de stocks variable, dépendant de la conjoncture du moment ? Dans le schéma présenté ici, les mouvements dus aux fluctuations du climat général traduiraient le comportement de stockage lié aux aléas conjoncturels proprement dits : en phase ascendante, pour des raisons tenant à la fois à une activité passée et à des anticipations de demande favorables, le niveau des stocks a tendance à être jugé de plus en plus léger et les industriels sont donc plutôt enclins à reconstituer leurs stocks, alors que les variations inverses se produisent en phase descendante. Dans ces conditions, l'information spécifiquement apportée par le solde d'opinion relatif à la question sur les stocks fournirait un renseignement sur l'écart de comportement par rapport à ce scénario de référence.

Les fluctuations de l'information spécifiquement apportée par la réponse sur les **perspectives générales de production** (cf. *graphique 6* en annexe) sont amples et persistantes. La remarque faite à propos de la question sur les perspectives personnelles de production demeure valable : il semble que le pouvoir prédictif de cette information se soit amoindri depuis le début des années 80, de sorte qu'en faisant passer le solde d'opinion au dessus ou en dessous du climat général qui se dégage de l'enquête, elle constitue sans doute plus, à présent, une sorte de miroir du discours ambiant sur l'état de la conjoncture, et de ses excès éventuels.

3.5 Le mélange des enquêtes mensuelles et trimestrielles

Comme on l'a vu dans la partie 2.3, la représentation espace-état permet facilement de mélanger les soldes d'opinion relatifs au même secteur (l'industrie), mais ayant des périodicités différentes. L'enquête trimestrielle fournit (entre autres) les

réponses à neufs questions : les opinions relatives à la demande globale ou étrangère passée ou future, celles concernant les effectifs passés et futurs, le taux d'utilisation des capacités de production, le jugement sur ces capacités de production, enfin l'opinion sur les goulots de production. Les données sont disponibles depuis le premier trimestre de 1976. Nous estimons un modèle à composantes inobservables par filtre de Kalman, en utilisant ces neufs soldes trimestriels plus les six soldes mensuels précédents, sur la période 1976:T1-1996:T1.

Les résultats obtenus confirment le lien entre les deux enquêtes, car le facteur commun est extrêment proche de l'indicateur précédent (calculé à partir de la seule enquête mensuelle) ; il explique au moins 50% de la variance de chacun des soldes d'opinion. Les paramètres gouvernant la dynamique de l'indicateur sont quasiment identiques à ceux obtenus lors de la précédente estimation (et ce, bien que la période d'estimation ne soit pas la même) :

$$F_t = \underset{(0.05)}{1.88}\, F_{t-1} - \underset{(0.05)}{0.90}\, F_{t-2} + u_t - \underset{(0.11)}{0.61}\, u_{t-1}$$

De plus, ce « nouvel »indicateur de climat est un peu plus lisse que le précédent, puisque le ratio de la variance de l'innovation sur la variance de l'indicateur est inférieur de 14% ; ceci n'est d'ailleurs pas réellement surprenant, l'estimation tenant compte d'une quantité plus importante d'information. Enfin, les composantes spécifiques obtenues peuvent là encore recevoir une interprétation, exercice auquel nous ne nous livrons pas ici car il serait quelque peu fastidieux. Contentons-nous de remarquer que les questions relatives aux effectifs ou aux capacités de production fournissent des informations spécifiques particulièrement utiles s'agissant des évolutions possible de l'emploi ou de l'investissement.

4 Étude à un niveau plus désagrégé

4.1 Analyse factorielle : cas de plusieurs facteurs communs

Comme nous l'avons mentionné précédemment, lorsque plusieurs facteurs sont nécessaires pour rendre compte des données étudiées, les facteurs initialement calculés par la procédure ne sont pas toujours aisément interprétables. Il est souvent nécessaire de les soumettre à une rotation, afin d'obtenir de nouveaux facteurs ayant davantage de signification économique. La rotation en question peut même être une simple transformation linéaire (rotation oblique), c'est-à-dire que l'hypothèse de non-corrélation des facteurs peut être abandonnée.

Bien sûr, le mot "interprétable" contient une part de jugement subjectif. Mais il est cependant possible de donner un contenu statistique à ce mot. En effet, plus les facteurs représentent un groupe donné de variables, plus ils sont faciles à interpréter. Il est clair, par ailleurs, qu'un facteur F_j représente un groupe de variables $y_{i_1}, \ldots, y_{i_p}$, si deux conditions sont vérifiées :

- pour $r = 1, \cdots p$, les pondérations $\lambda_{i_r,j}$ sont nettement plus importants que les pondérations $\lambda_{i_r,k}$ correspondant à l'un quelconque des autres facteurs F_k ;

- ils sont aussi beaucoup plus importants que les pondérations λ_{ij} associés à l'une quelconque des autres variables y_i, pour $i \notin \{i_1, \ldots i_k\}$.

En pratique, toutes les procédures existantes consistent à effectuer une rotation des facteurs initiaux de façon à ce que les facteurs obtenus à l'issue de cette rotation soient associés au plus grand nombre possible de pondérations nulles (ou, au moins négligeables). Si l'on note F_j^* les facteurs obtenus après rotation, le modèle devient :

$$y_{it} = \lambda_{i1}^* F_{1t}^* + \ldots + \lambda_{iJ}^* F_{Jt}^* + u_{it},$$

dans lequel $\Lambda^* = \left(\lambda_{ij}^*\right)_{\substack{1 \le i \le I \\ 1 \le j \le J}}$ est une matrice contenant de nombreux termes nuls ou voisins de zéro, et les facteurs F_{jt}^* sont généralement corrélés (mais restent, bien sûr, non corrélés avec les u_{it}).

Dans les logiciels d'analyse factorielle, plusieurs procédures sont généralement disponibles pour effectuer de telles rotations - sans rentrer dans le détail de ces procédures, précisons simplement qu'elles reposent le plus souvent sur la maximisation d'un critère quadratique. Nous avons, pour notre part, utilisé la méthode PROMAX (voir Cureton et Mulaik, 1975), disponible dans le logiciel SAS.

4.2 Application à l'enquête mensuelle au niveau des grandes branches industrielles

L'analyse factorielle a montré qu'une part importante de l'information contenue par les réponses de l'enquête mensuelle dans l'industrie pouvait être résumée par le comportement d'une seule variable cachée, que nous avons appelé « climat général des affaires ». Néanmoins, une telle dénomination n'est réellement justifiée que si ce

climat se trouve présent à un niveau plus désagrégé, parmi les branches du secteur industriel : si cet indicateur ne constituait, en définitive, que la résultante de conjonctures sectorielles très indépendantes les unes des autres, une analyse menée au niveau le plus global se révélerait plus formelle que véritablement pertinente : parler d'un climat général des affaires aurait, dans ces conditions, moins d'intérêt. Mieux vaudrait se concentrer directement sur les variations observées au niveau des branches.

C'est pourquoi il semble nécessaire d'appronfondir quelque peu la nature de l'indicateur de climat en modélisant ensemble les variations de cinq soldes d'opinion [6] pris au niveau de trois grands secteurs, à savoir celui des biens intermédiaires, celui des biens d'équipement et celui des biens de consommation. La période d'estimation est mars 1976-janvier 1996.

L'analyse factorielle offre un critère pour déterminer le nombre de facteurs communs, en calculant les valeurs propres successives de la matrice $\Lambda\Lambda'$. Les résultats obtenus ainsi que la proportion des communautés totales expliquée par chaque valeur propre sont :

	valeur propre	proportion
1	91.81	79.8%
2	12.02	10.6%
3	7.27	6.4%
4	2.15	1.9%
5	1.34	0.67%
...	...	...

Deux enseignements peuvent être tirés de ces résultats. D'une part, on est amené à retenir trois facteurs communs. D'autre part, le premier de ces trois facteurs est largement prépondérant.

Lorsque l'on compare les variations de ce premier facteur avec l'indicateur de climat des affaires précédemment obtenu (cf *graphique 5* en annexe), les deux variables se révèlent extrêmement proches l'une de l'autre. Ceci signifie précisément que l'indicateur de climat, qui a été estimé à l'aide des soldes d'opinion

6 Ces soldes d'opinion concernent les niveaux de production passé et futur, les niveaux des carnets globaux et des carnets étrangers, enfin le niveau des stocks. Les réponses relatives aux perspectives générales d'activité ne sont pas disponibles par branche.

au niveau le plus agrégé, est « réellement » et fortement présent au sein des trois branches étudiées ; ce qui justifie *a posteriori* la dénomination que nous lui avons donnée.

Néanmoins, cet indicateur de climat ne suffit pas pour rendre compte de l'information commune contenue dans les quinze soldes d'opinion : d'autres liens existent lorsque l'analyse est menée à un niveau plus fin. D'où le problème d'interpréter les deux autres facteurs. Pour ce faire, on est amené à appliquer une transformation linéaire aux facteurs, de façon à ce que les nouveaux facteurs ainsi obtenus puissent faire l'objet d'une interprétation. Dans le cas présent, la recombinaison des trois facteurs communs initiaux fournit des résultats facilement explicables : une fois appliquée la procédure PROMAX, il apparaît clairement que les nouveaux facteurs obtenus sont relatifs à chacune des branches (cf *graphique 7* en annexe). En effet, lorsque l'on mène séparément sur chaque secteur une analyse factorielle, le facteur commun obtenu se trouve très proche de l'un des trois facteurs fournis par le modèle d'ensemble.

Pour résumer ce qui vient d'être dit, il apparaît en définitive que la construction d'un indicateur de climat pour l'industrie prise dans son ensemble comme pour chaque grande branche a du sens : le climat des affaires de l'industrie est présent à un niveau plus désagrégé, mais il se décline suivant des modalités différentes. En d'autres termes, on est fondé à parler d'un climat pour chaque grande branche, dont les variations diffèrent certes d'une branche à l'autre, mais présentent tout de même une grande cohérence : c'est cette cohérence dont rend compte le climat d'ensemble.

4.3 Interprétation des facteurs communs et spécifiques au niveau des branches

Nous présentons maintenant les résultats obtenus dans les différentes branches, en commentant tout d'abord les climats de chacune d'entre elles, avant d'en venir à la composante spécifique à chaque question.

Si l'on en croit les évolutions comparées des climats branche par branche (cf *graphique 6* en annexe), les effets du second choc pétrolier ont été moins prononcés dans la branche des biens d'équipement que dans les deux autres branches. Les industries de biens de consommation ont, très logiquement, bénéficié d'un rebond plus important que celles des biens intermédiaires lors du plan de relance gouvernemental de 1981. Ces dernières, sans doute plus sensibles à la conjoncture internationale, ont en revanche été tirées plus tôt par la reprise mondiale du milieu des années 80 (dès 1983 aux Etats-Unis, en 1984 en RFA), et semblent en avance sur les autres branches lors du cycle d'expansion initié par le contre-choc pétrolier, mais aussi au moment du ralentissement qui a suivi.

La décélération en deux temps (ralentissement en 90-91, fausse reprise en 92 et récession profonde en 93) est visible dans les trois secteurs, mais on peut noter que les biens d'équipement avaient touché un point bas dès 1991 (le niveau de l'indicateur correspondant est inférieur dès cette date à celui du début des années 80, ce qui n'est pas le cas pour les autres branches), ce qui constitue une illustration du recul de l'investissement trois années durant. La reprise de 1994 semble simultanée dans les trois branches, mais son ampleur apparaît plus marquée dans le secteur des biens intermédiaires : le retournement provenait pour une grande part d'un mouvement sur les stocks, qui stimula le commerce interindustriel.

L'analyse de la composante spécifique de chaque réponse recueillie au niveau des branches est quelquefois rendue difficile par le profil instable que certaines d'entre elles peuvent avoir (instabilité en général plus heurtée qu'au niveau de l'industrie dans son ensemble). Néanmoins, l'étude systématique des questions pour lesquelles l'information spécifique est importante peut aider le conjoncturiste dans la formulation de son diagnostic. Plutôt que de se livrer à un commentaire conjoncturel des quinze soldes d'opinion (cinq questions, trois branches), on se contentera ici de dégager quelques faits saillants.

Au niveau de l'industrie prise dans son ensemble, la composante commune semblait très proche des de la réponse portant sur la **tendance passée de la production** (cf *graphique 8.1* en annexe) et, dans une moindre mesure, de celle portant sur le **niveau des carnets de commande** (cf *graphique 8.3* en annexe). La branche des biens d'équipement semble toutefois sortir nettement de ce schéma : les carnets de commande y jouent un rôle beaucoup plus important que la tendance passée de la production. La longueur et le coût des processus de production, sans doute plus importants pour ce qui concerne les biens d'équipement, peuvent expliquer que la réponse concernant la tendance passée de la production y soit nettement moins proche de l'indicateur de climat que dans les autres secteurs : la conjoncture y apparaît bien plus dépendante du nombre de commandes d'ores et déjà enregistrées, et donc de l'état des carnets, que de la production effectuée dans les mois écoulés.

Cette particularité explique peut-être aussi la spécificité de la question relative aux **stocks dans les industries de biens d'équipement** (cf *graphique 8.5* en annexe). C'est en effet la seule question qui soit "mal expliquée" (en terme de décomposition de variance) par la composante commune : le niveau des stocks est sans doute bien moins rapidement adaptable aux aléas de la conjoncture qu'il ne l'est dans les deux autres secteurs.

Au-delà d'un intérêt d'ordre conjoncturel, l'amplitude moyenne de la composante spécifique relative à la question sur **la demande et les carnets étrangers** (cf *graphique 8.4* en annexe) constitue une mesure indirecte du poids des marchés étrangers dans la détermination de l'activité du secteur. On constate ainsi sur longue période que la branche pour laquelle ce solde d'opinion est le plus proche du climat

sectoriel correspondantest celle des biens d'équipement, suivie par celles des biens intermédiaires. Cette plus grande influence de l'étranger sur l'activité est conforme avec les données de la comptabilité nationale, puisque la part de la production exportée se montait en 1994 à environ 45% dans la branche des biens d'équipement professionnels, à 37% dans celle des biens intermédiaires et à seulement 29% pour ce qui concerne les biens de consommation.

Conclusion

Dans cette étude, nous estimons des modèles factoriels dynamiques en utilisant aussi bien des techniques standard d'analyse factorielle que le filtre de Kalman. Nous montrons que ce type de modèle peut être utilisé pour aider à analyser l'information contenue dans les enquêtes de conjoncture.

Une extension naturelle de ce travail consisterait à mélanger des données provenant de différents secteurs de façon à construire un indicateur de climat plus général. La modélisation espace-état permet aisément de résoudre la difficulté technique relative au mélange de données de périodicités différentes.

Une autre extension possible serait d'utiliser le même genre de techniques à des fins prévisionnelles. En fait, l'indicateur construit ici semble plutôt coïncident. Une étape supplémentaire consisterait à chercher à construire un indicateur avancé d'activité.

BIBLIOGRAPHIE

Cureton E .E., Mulaik S. A. (1975). « The weighted Varimax rotation and the Promax rotation », *Psychometrika*, 40, 183-195

Engle R.F., Watson M.W., (1981). « A one factor multivariate time series model of metropolitan wage rates », *Journal of American Statistical Association*, 76, 774-781.

Engle R.F., Watson M.W., (1983). « Alternative algorithms for the estimation of dynamic factor, MIMIC, and varying coefficient regression models », *Journal of Econometrics*, 23, 385-400.

Forni, M., Reichlin L. (1995). « Let's get real : a dynamic factor analytical approach to disaggregated business cycle », *Center for Economic Policy Research Discussion Paper n°1244.*

Geweke J. (1977). « Labor turnover and employment dynamics in US manufacturing », in *New methods in business cycles research, Sims Ed.*, Minneapolis : Federal Reserve Bank of Minneapolis.

Geweke J., Singleton K.J. (1981). « Maximum likelihood "confirmatory" factor analysis of economic time series », *International Economic Review*, 22 (1), 37-54.

Gourieroux C., Monfort A. (1990) *Séries temporelles et modèles dynamiques*, Economica

Harvey, A.C. (1989). *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press.

Lawley, D.N., Maxwell A.E. (1971). *Factor Analysis as a Statistical Method,* New York Macmillan Publishing Co.,Inc.

Newey W.K., Mac Fadden D. (1994). « Large Sample Estimation and Hypothesis testing », *in Handbook of Econometrics,vol.4,* 2113-2245.

Quah D., Sargent T.J. (1993). « A dynamic index model for large cross-sections », in *Business cycles, indicators and forecasting, J.H.Stock and M.W.Watson Ed,* University of Chicago Press.

Sargent, T.J., Sims S.A. (1977). « Business cycle modelling without pretending to have too much *a priori* economic theory », in *New methods in business cycles research, Sims Ed.*, Minneapolis : Federal Reserve Bank of Minneapolis.

Stock, J.H., Watson M.W. (1989). « New indexes of coincident and leading indicators », in *NBER Macroeconomics Annual, Blanchard & Fisher Ed.*, MIT Press, Cambridge.

Stock J.H., Watson M.W. (1991). « A probability model of the coincident economic indicators », in *Leading economic indicators : new approaches and forecasting records, K.Lahiri and G.H.Moore Ed,* Cambridge University Press.

Stock J.H., Watson M.W. (1993). « A procedure for predicting recessions with leading indicators : econometric issues and recent experience », in *Business cycles, indicators and forecasting, J.H.Stock and M.W.Watson Ed,* University of Chicago Press.

Watson M.W., Kraft D.F. (1984). « Testing the interpretation of indices in a macroeconomic index model », *Journal of Monetary Economics*, 13, p 165-181

Watson M. W. (1986). « Univariate detrending methods with stochastic trends », *Journal of Monetary Economics,* 18, 49-75

Preuve du lemme 1:

i) Il s'agit de vérifier que si l'un des paramètres λ_{ij}, $i = 1...I$, $j = 1,...J$ ou σ_i^2, $i = 1...I$ tend vers l'infini alors $Q_T(y,\theta) \to -\infty$. Nous allons en effet montrer que, dans ce cas, $det\, C \to +\infty$, alors que $tr\, C^{-1}A$ reste borné, ce qui entraîne le résultat annoncé.

- Montrons d'abord que : si $\|C\| \to +\infty$, alors $det\, C \to +\infty$.

Comme on a supposé que $\sigma_i^2 \neq 0$, pour $i = 1...I$, la matrice D est définie positive. Comme, en outre, $C \geq D$, C est aussi définie positive. Pour montrer que $det\, C \to +\infty$, il suffit donc de montrer que l'une au-moins des valeurs propres de C tend vers $+\infty$.

Or on sait que : $\|M\|_2^2 = \underset{y \neq 0}{Max} \dfrac{y'M'My}{y'y}$ définit une autre norme sur l'espace des matrices, et que, si l'on désigne par α_i, $i = 1...I$ les valeurs propres de C matrice définie positive, cette norme vérifie : $\|C\|_2 = \underset{i}{Max}\, \alpha_i$.

Par équivalence des normes sur l'espace des matrices, on sait que $\|C\|_2 \to +\infty$, lorsque $\|C\| \to +\infty$. On obtient ainsi le résultat.

- Montrons maintenant que $tr\, C^{-1}A$ reste borné. Tout d'abord, il est clair que :

$$\left| tr\, C^{-1}A \right| \leq I^2 \left\| C^{-1} \right\| \|A\| .$$

Si l'on note : $\sigma^2 = \underset{i}{Min}\, \sigma_i^2$, comme $C \geq D > 0$, on sait que les valeurs propres de C sont toutes supérieures à σ^2. D'après le résultat précédemment cité sur l'équivalence des normes de matrices, il existe donc un réel k tel que : $\left\| C^{-1} \right\| \leq \dfrac{k}{\sigma^2}$. Comme on a, de plus, supposé qu'il existe un réel $\alpha > 0$, tel que : $\forall \theta \in \Theta, \forall i = 1,... I, \sigma_i^2 \geq \alpha$, on obtient : $\left| tr\, C^{-1}A \right| \leq \dfrac{kI^2}{\alpha} \|A\|$, d'où le résultat.

Preuve du lemme 2 : Il faut montrer que : $\underset{\theta\in\Theta}{Max}\left|Q_T(y,\theta)-Q_0(\theta)\right|\underset{P}{\to} 0$.

Or cette expression est égale à : $\underset{\theta\in\Theta}{Max}\left|tr\,C^{-1}\left(\frac{1}{T}\sum_t x_t\,x_t'\right)-trC^{-1}C_0\right|$, et on peut écrire :

$$\left|tr\,C^{-1}\left(\frac{1}{T}\sum_t x_t\,x_t'\right)-trC^{-1}C_0\right|\le\left|tr\,C^{-1}\left(\frac{1}{T}\sum_t x_t\,x_t'-C_0\right)\right|\le I^2\left\|C^{-1}\right\|\left\|\frac{1}{T}\sum_t x_t\,x_t'-C_0\right\|$$

En raisonnant comme précédemment, on obtient :

$$\left|Q_T(y,\theta)-Q_0(\theta)\right|\le\frac{kI^2}{\alpha}\left\|\frac{1}{T}\sum_t x_t\,x_t'-C_0\right\|,$$

et cette inégalité est valable quel que soit $\theta\in\Theta$.

Comme par hypothèse (y_t) est stationnaire, et comme θ_0 est la vraie valeur du paramètre, on sait que : $\left(\frac{1}{T}\sum_t x_t\,x_t'\right)\underset{P}{\to}C_0$.

On déduit alors de ce qui précède que : $\left|Q_t(y_t,\theta)-Q_0(\theta)\right|\underset{P}{\to}0$, uniformément sur Θ.

Preuve du lemme 3 : On considère un processus (z_t) non autocorrélé, tel que pour tout t, la loi de z_t est la loi $N(0,C)$, où $C=\Lambda\Lambda'+D$ est définie comme précédemment, et où $C_0=\Lambda_0\Lambda_0'+D_0$ est la matrice de variance-covariance associée à la vraie valeur des paramètres. On note $f(z_t,\theta)$ et $f(z_t,\theta_0)$ les densités associées.

Les propriétés de l'information de Kullback conduisent à l'inégalité suivante :

$$E_{\theta_0}\left(Ln\left(\frac{f(z_t,\theta)}{f(z_t,\theta_0)}\right)\right)<0 \text{ si } \theta\neq\theta_0.$$

Pour $\theta\neq\theta_0$, cette inégalité entraîne que :

$$E_{\theta_0}\left(\frac{1}{T-1}\sum_t Ln\left(f(z_t,\theta)\right)\right) < E_{\theta_0}\left(\frac{1}{T-1}\sum_t Ln\left(f(z_t,\theta_0)\right)\right),$$

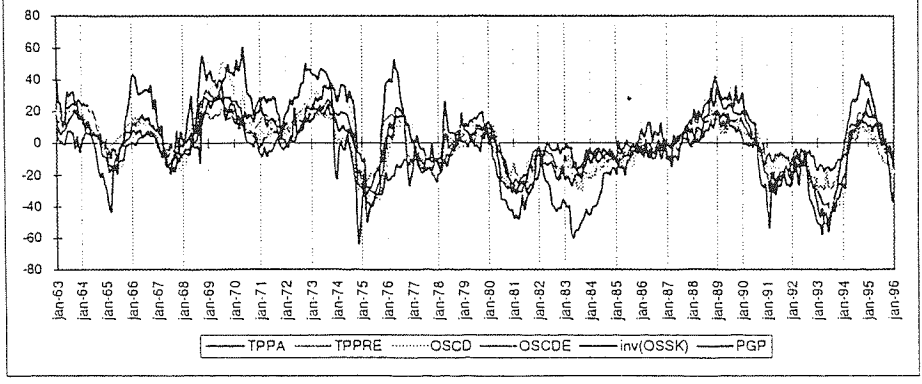
c'est-à-dire que : $-\frac{1}{2} Ln\ det\ C - \frac{1}{2} tr\ C^{-1}C_0 < -\frac{1}{2} Ln\ det\ C_0 - \frac{1}{2} tr\ C_0{}^{-1}C_0$,

ou encore que : $Q_0(\theta) < Q_0(\theta_0)$[7].

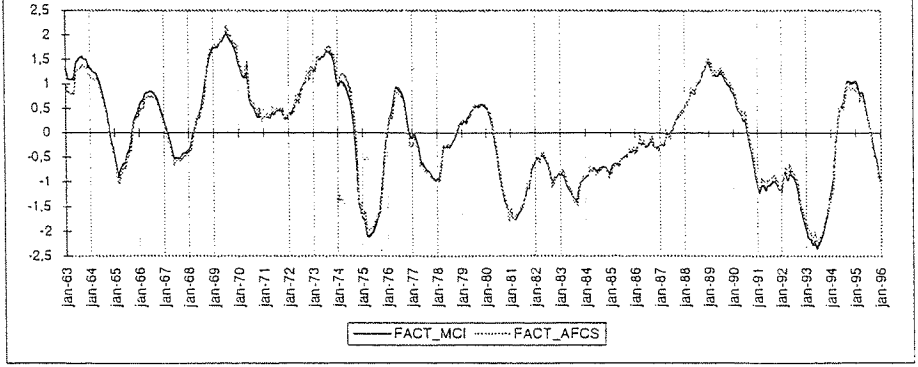
Preuve du lemme 4 : Ceci résulte de façon immédiate de la continuité de l'application qui à θ associe C, des applications *trace* et *déterminant*, ainsi que de l'application qui à C associe C^{-1} (cette dernière application étant bien définie sur le compact Θ).

7 On notera que le raisonnement qui est fait ici consiste à appliquer l'inégalité de l'information de Kullback à la pseudo-loi, en faisant momentanément comme si c'était la vraie loi. Cette inégalité est en effet une propriété de la loi étudiée, qu'elle soit ou non la vraie loi des observations.

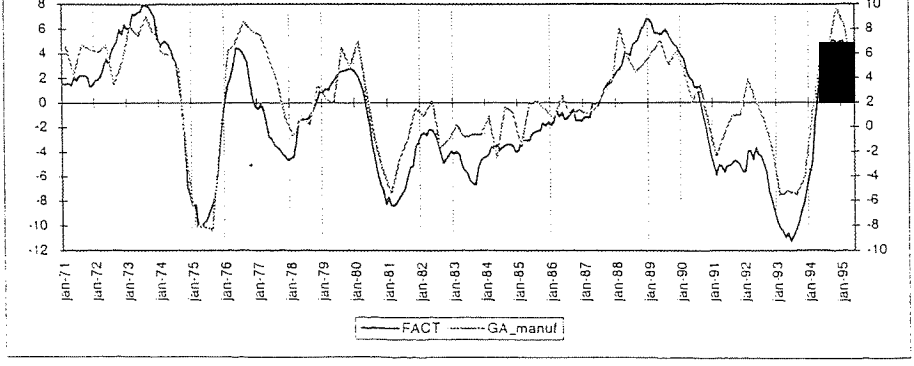
GRAPHIQUE 1
soldes d'opinions de l'enquête mensuelle



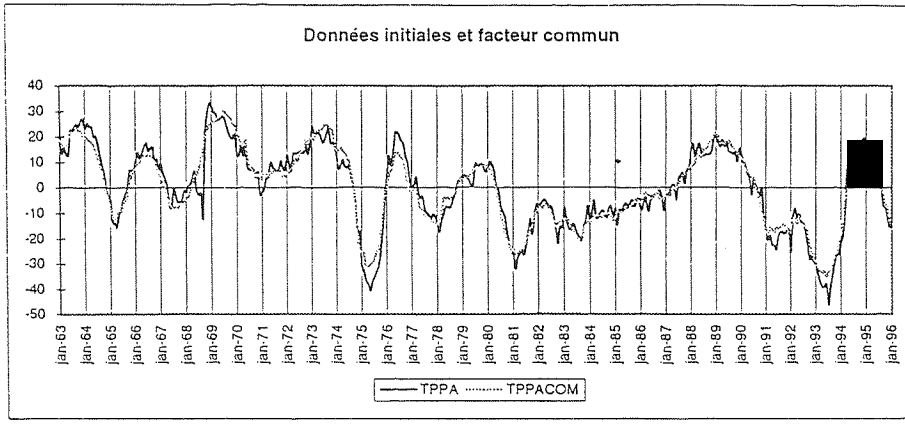
GRAPHIQUE 2 :
Comparaison des indicateurs obtenus par les deux méthodes

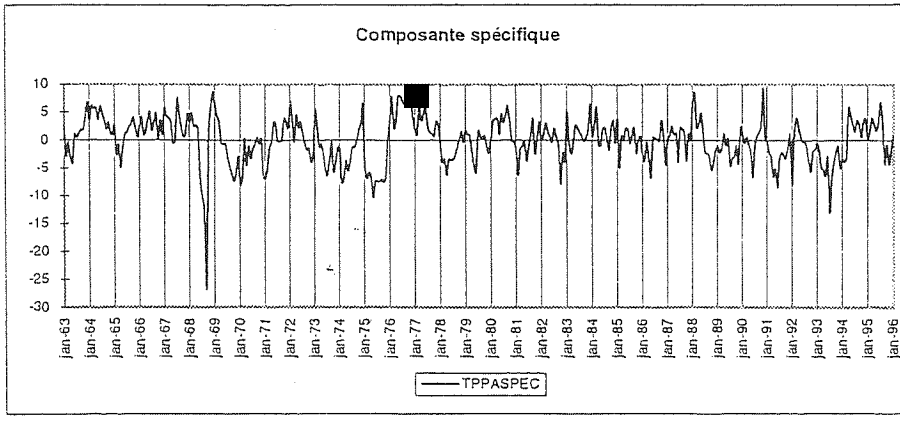


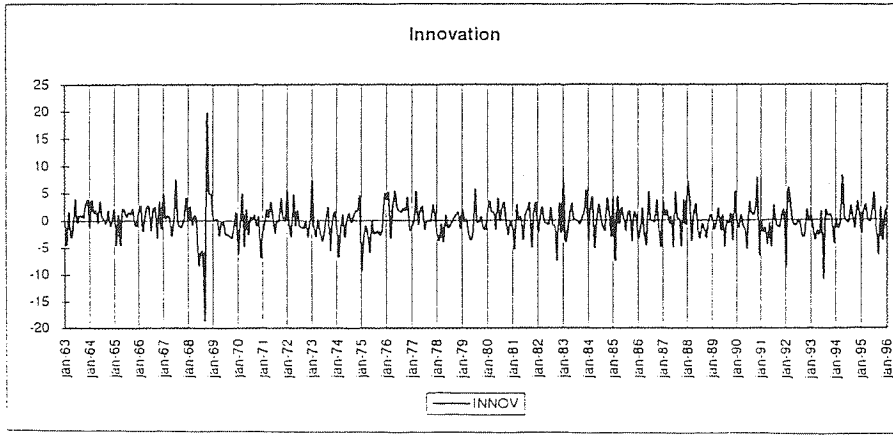
GRAPHIQUE 3
Indicateur composite et glissement annuel de la production manufacturière



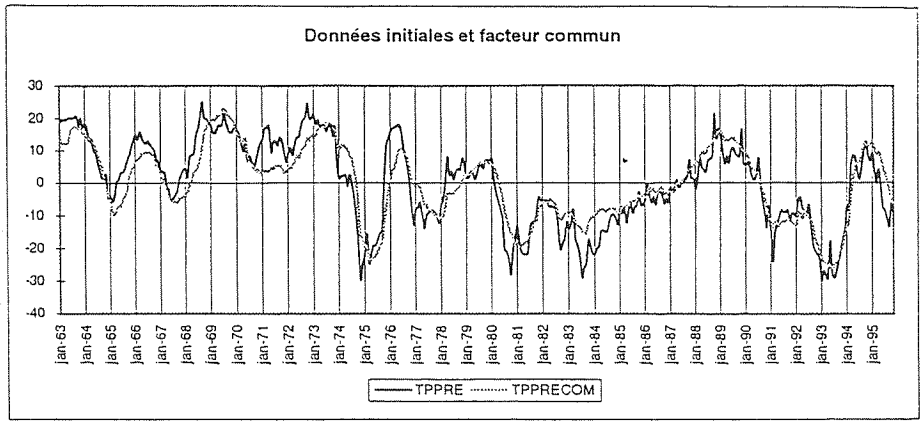
GRAPHIQUE 4.1
TENDANCE DE LA PRODUCTION PASSEE

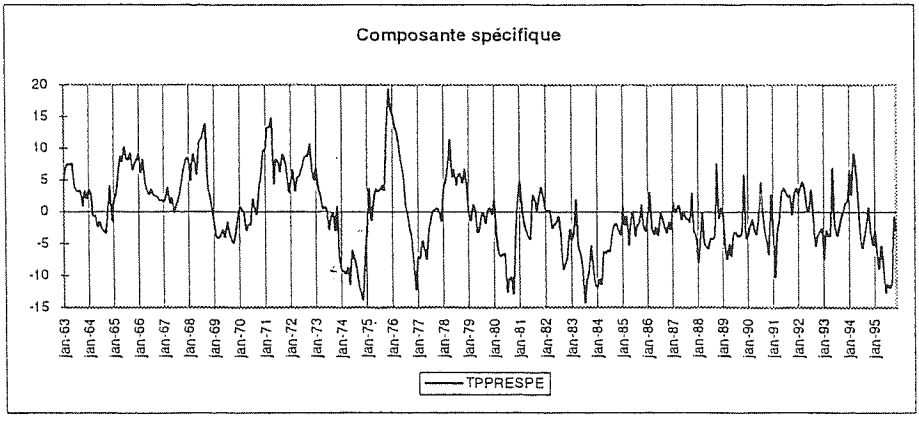


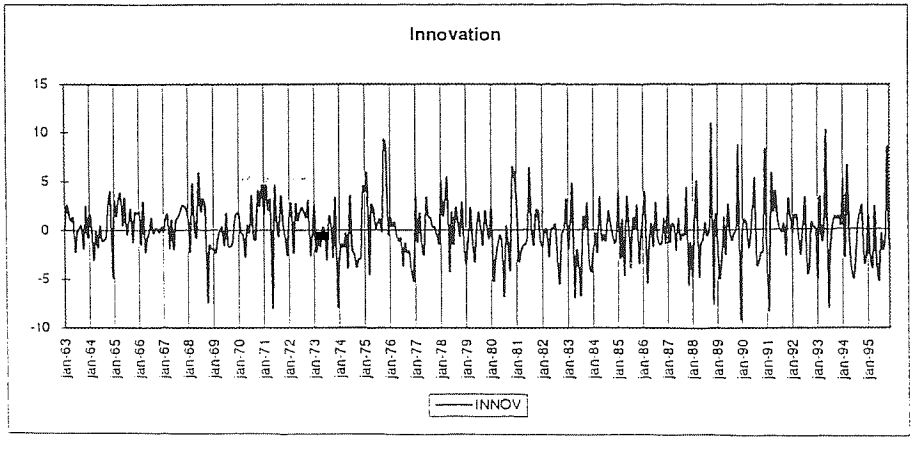




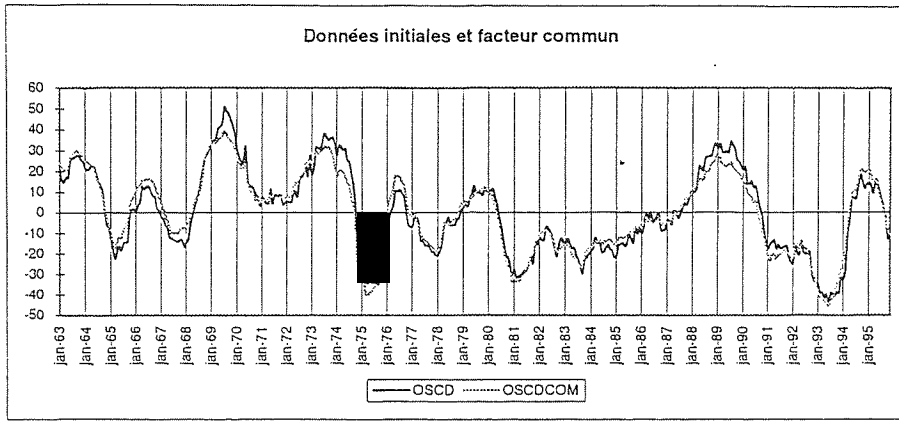
GRAPHIQUE 4.2
TENDANCE DE LA PRODUCTION PREVUE

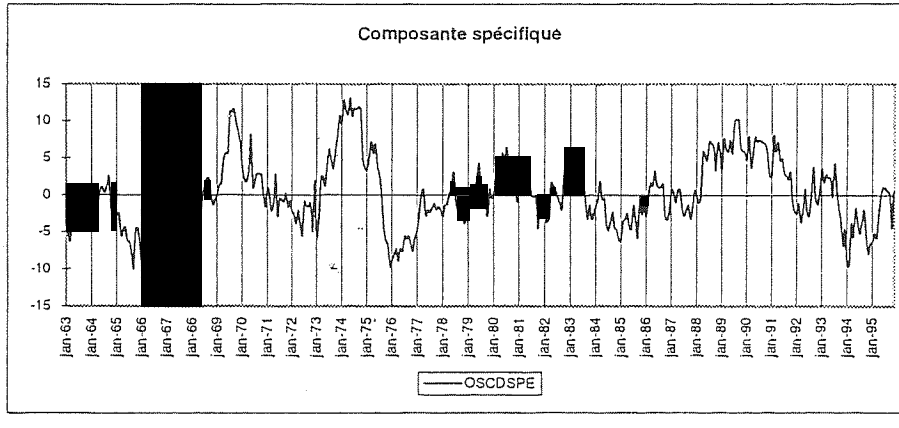


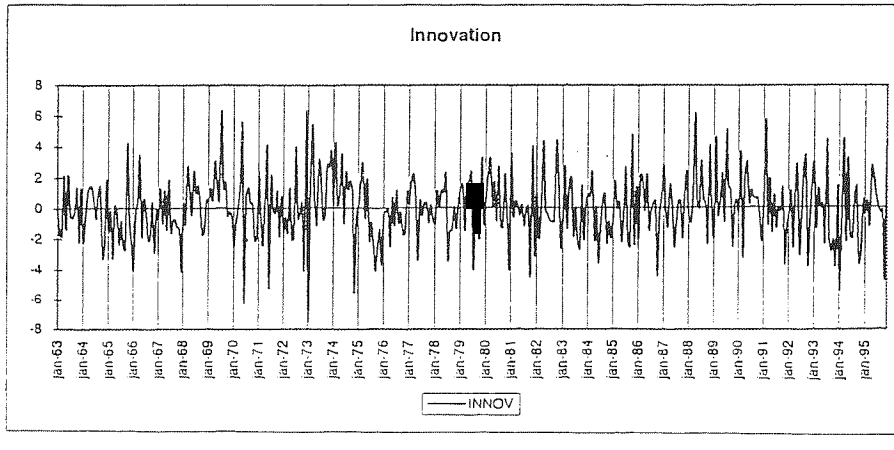




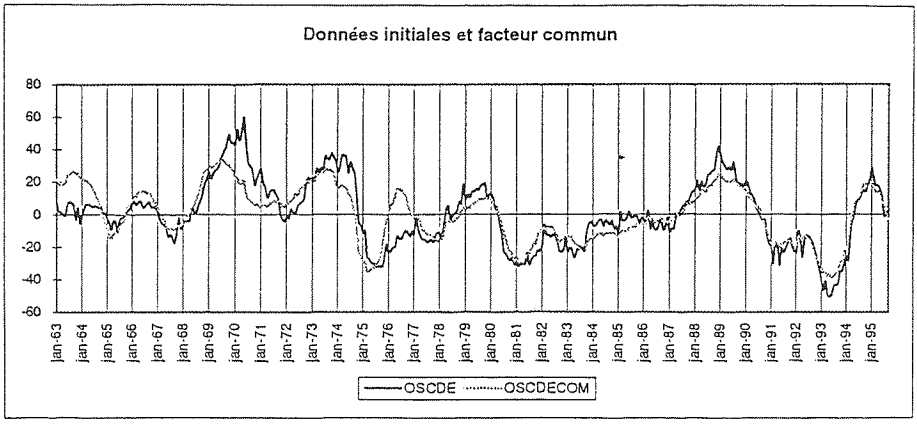
GRAPHIQUE 4.3
CARNETS DE COMMANDE

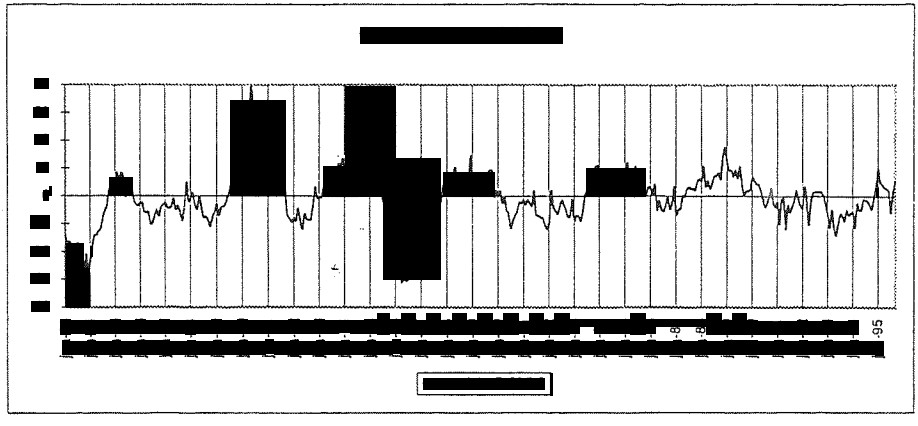


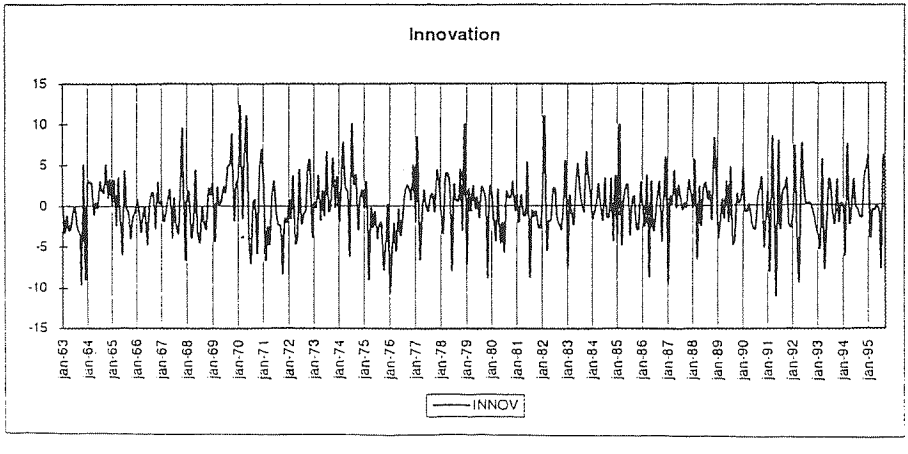




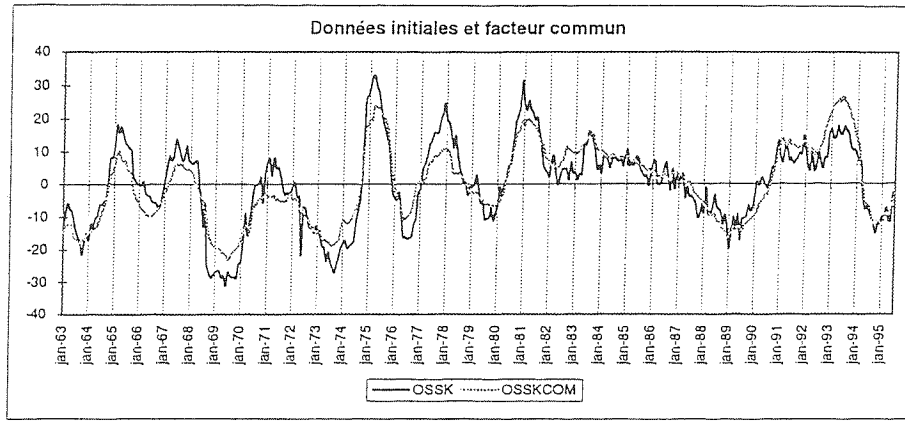
GRAPHIQUE 4.4
CARNETS DE COMMANDE ETRANGERS

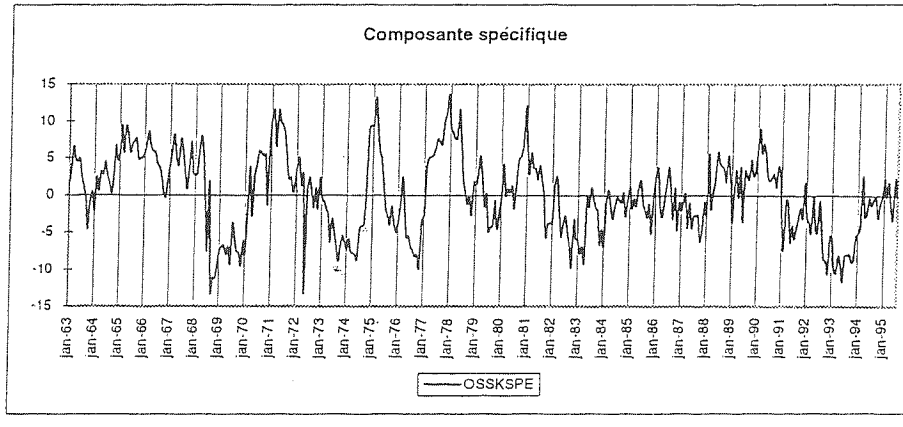


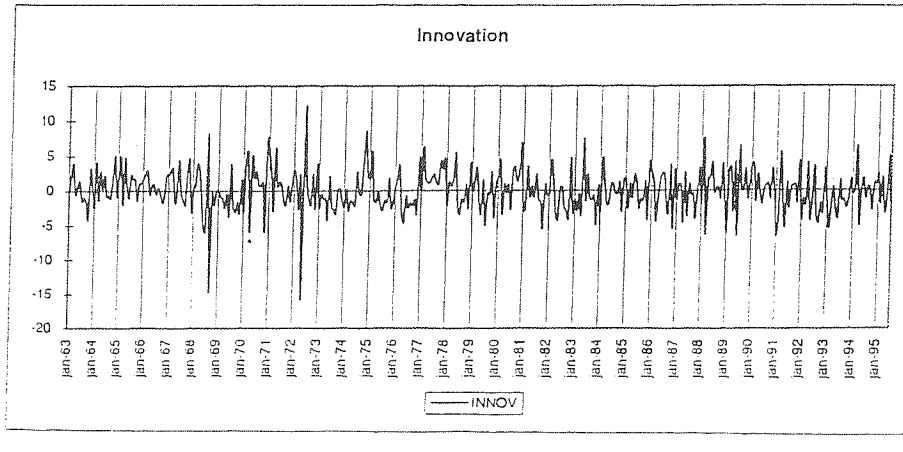




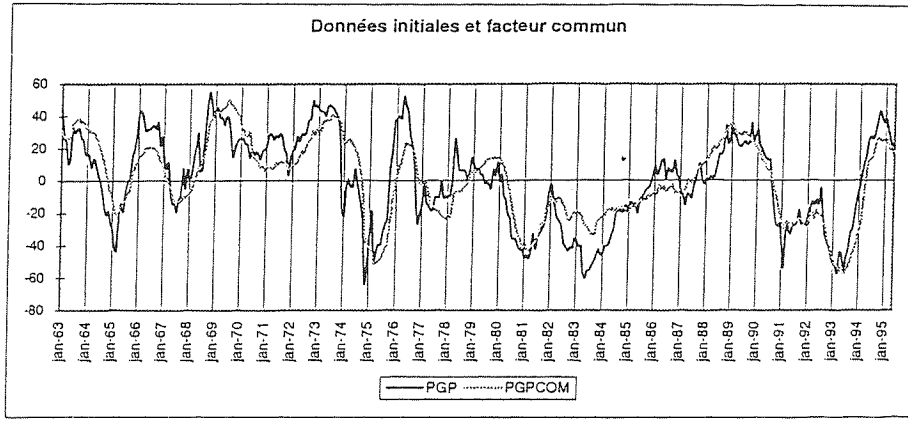
GRAPHIQUE 4.5
OPINION SUR LES STOCKS

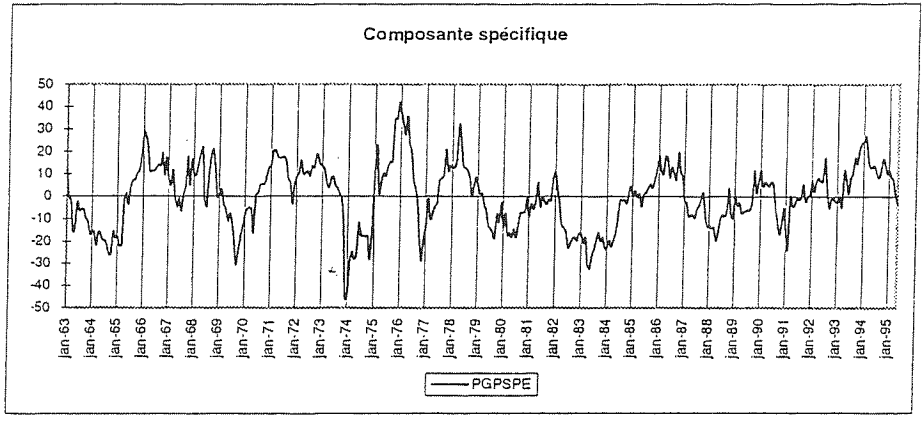


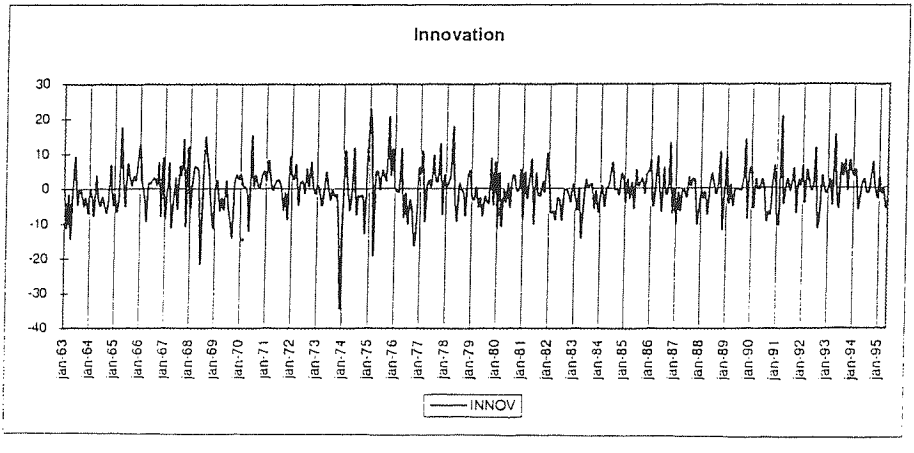




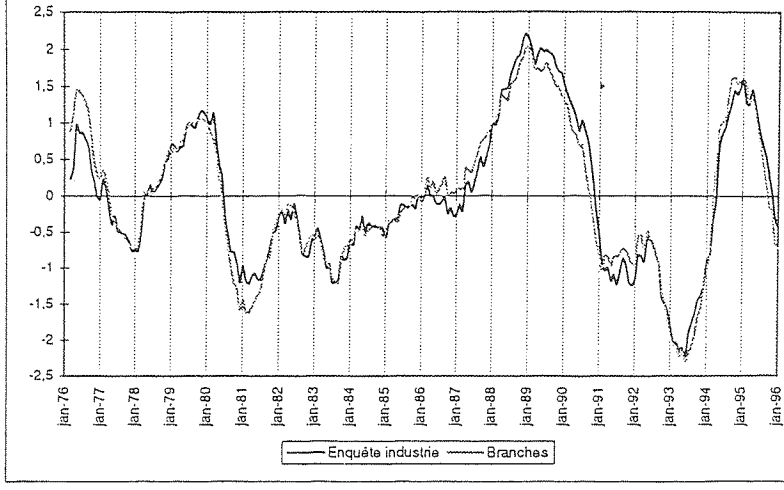
GRAPHIQUE 4.6
PERSPECTIVES GENERALES DE PRODUCTION



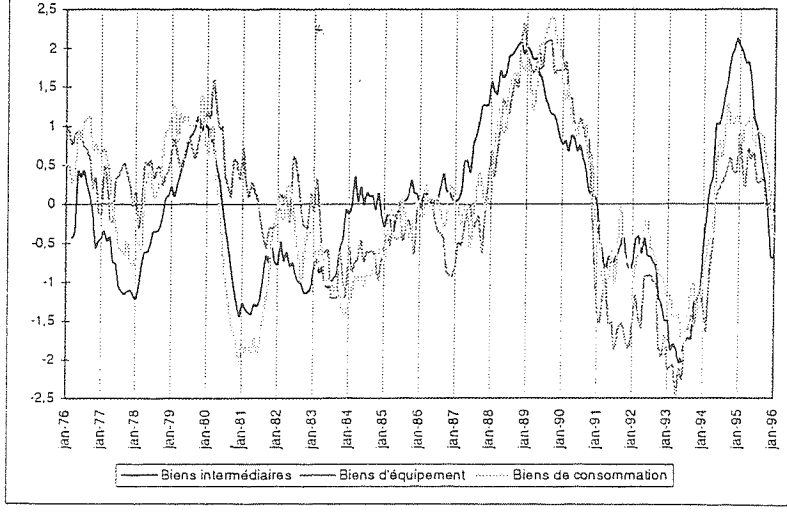




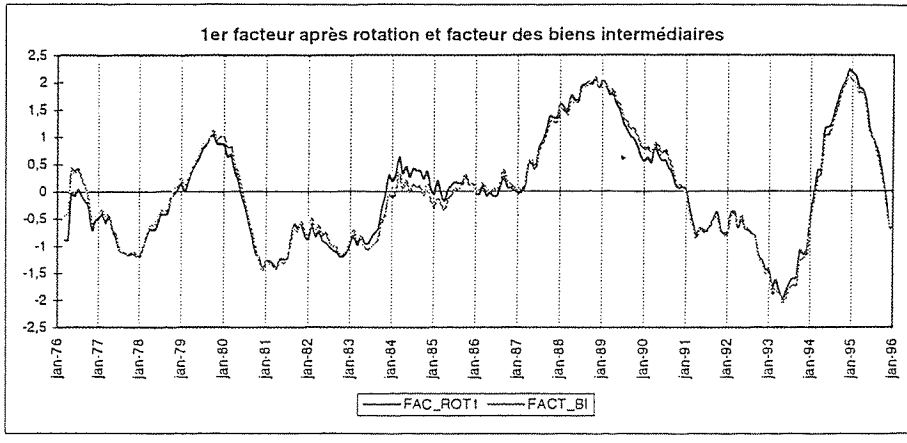
GRAPHIQUE 5
Facteur commun au niveau agrégé
et premier facteur commun au niveau branches

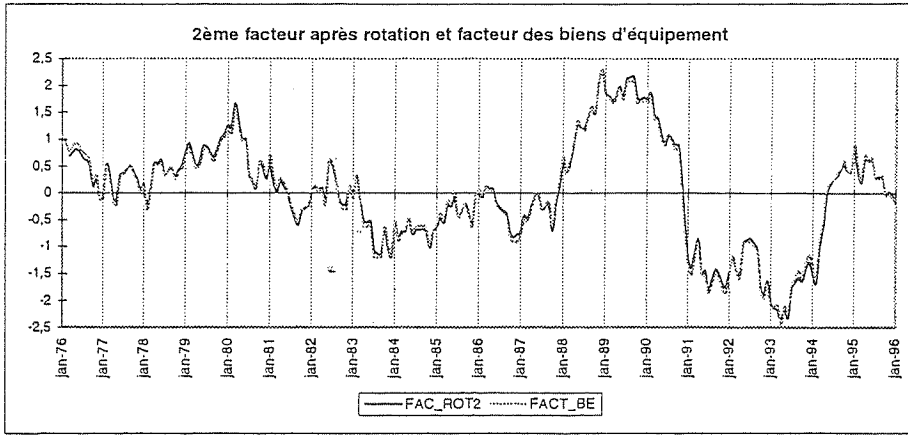


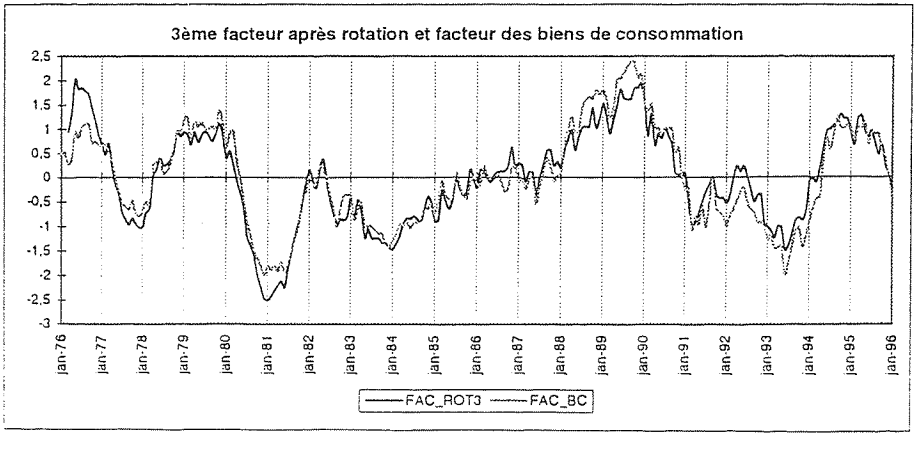
GRAPHIQUE 6
Comparaison des facteurs obtenus dans chaque branche



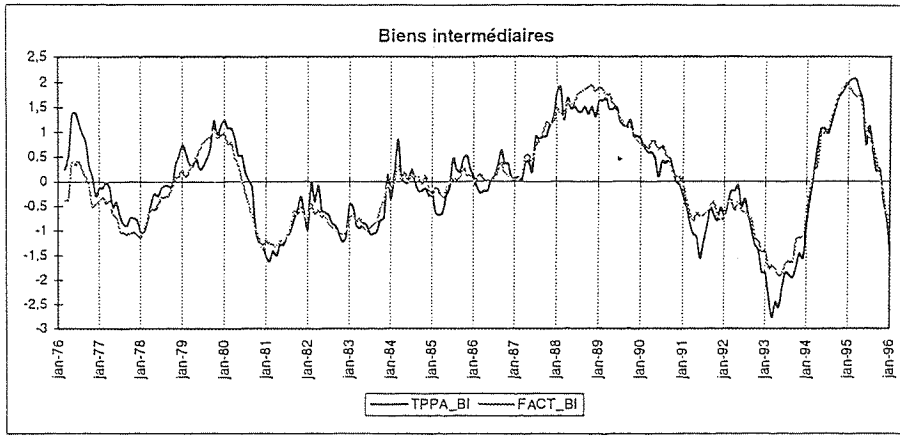
GRAPHIQUE 7
FACTEURS APRES ROTATION
ET FACTEURS OBTENUS DANS CHAQUE BRANCHE

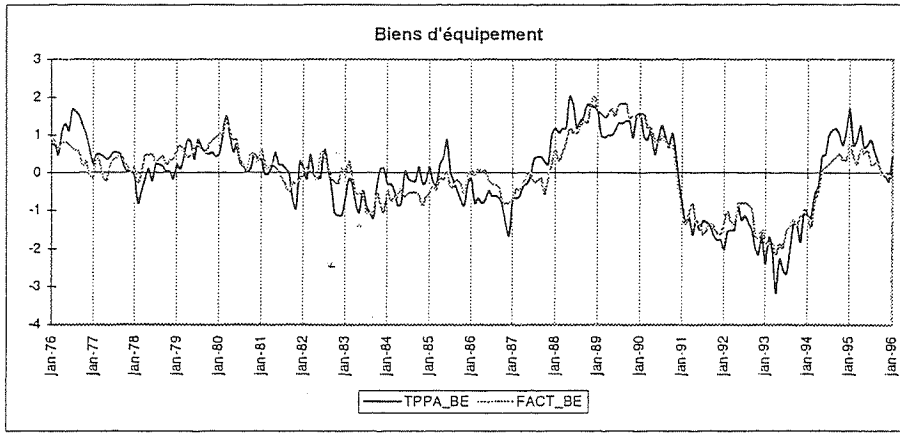


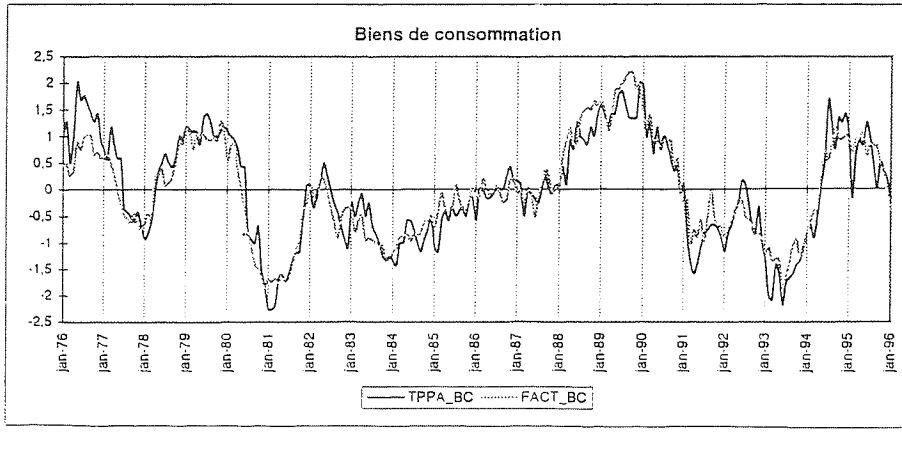




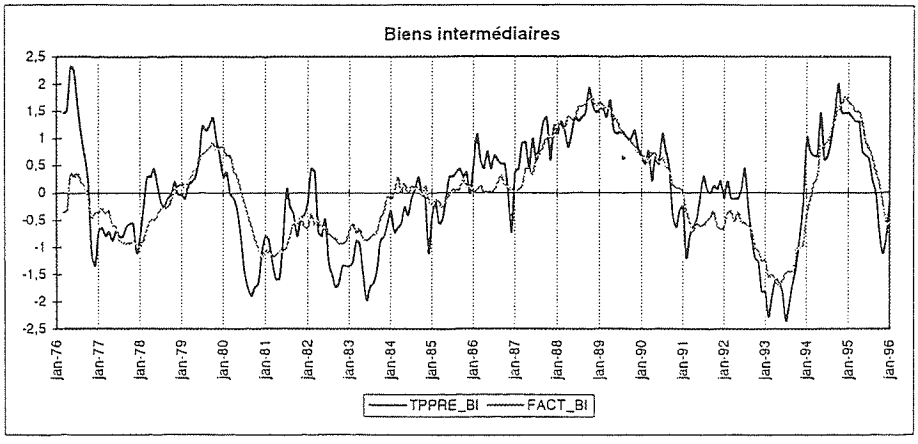
GRAPHIQUE 8.1 : TENDANCE DE LA PRODUCTION PASSEE
Données initiales et facteurs par branches

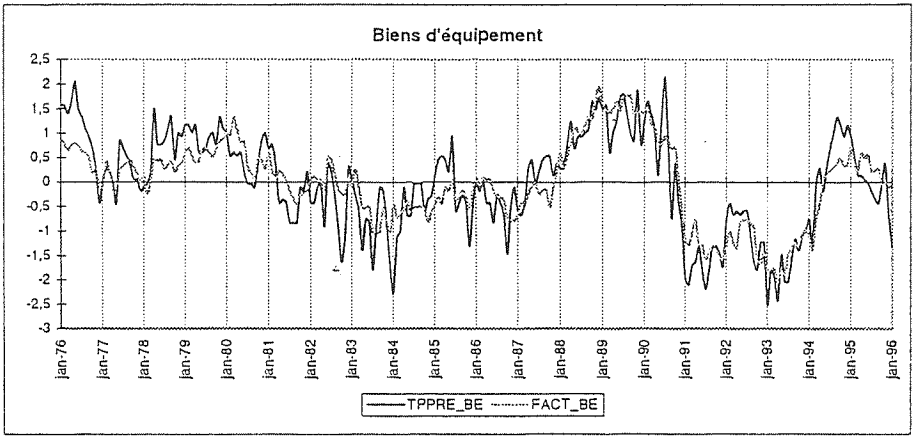


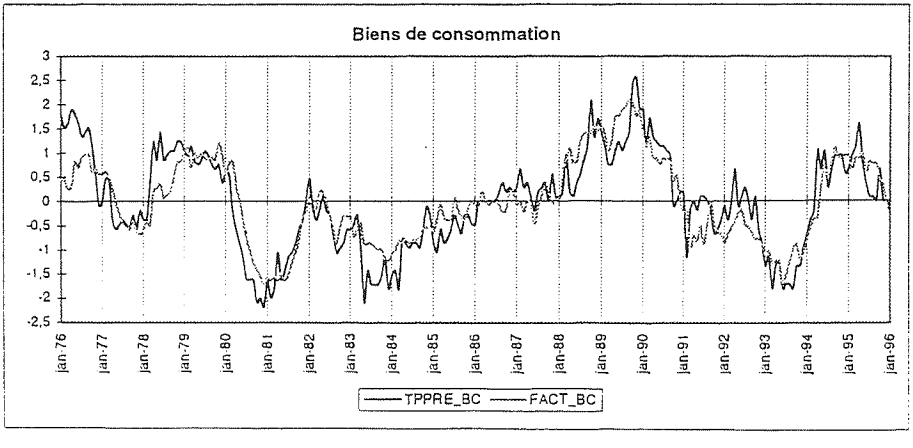




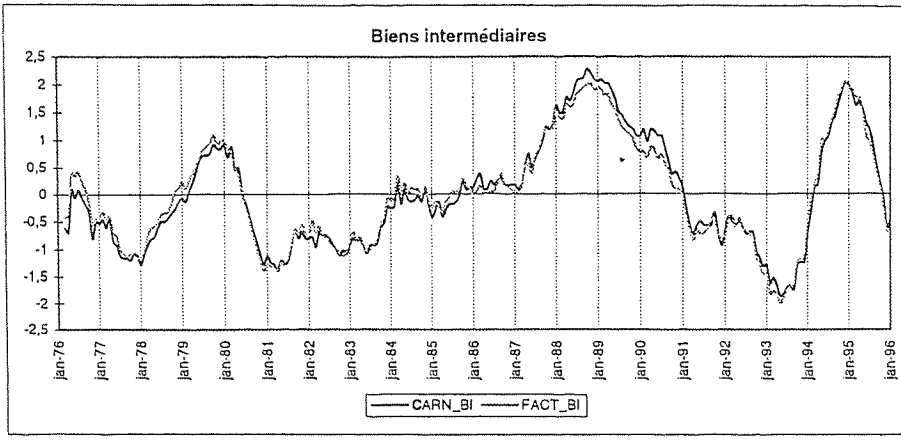
GRAPHIQUE 8.2 : TENDANCE FUTURE DE LA PRODUCTION
Données initiales et facteurs communs par branches

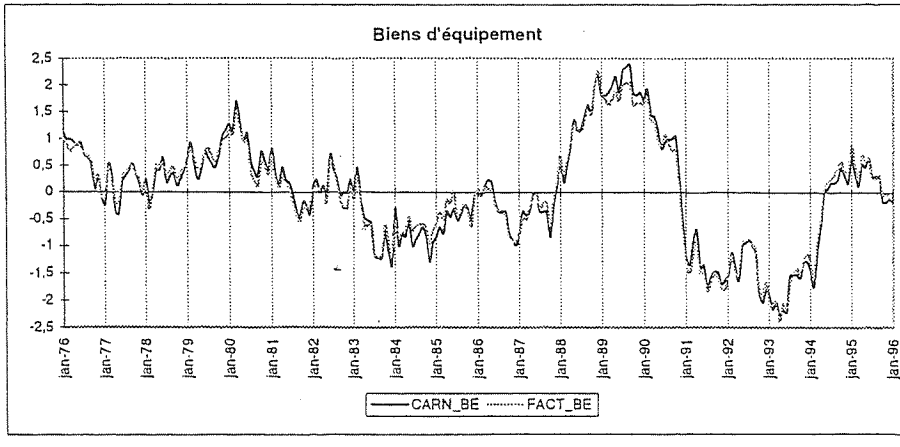


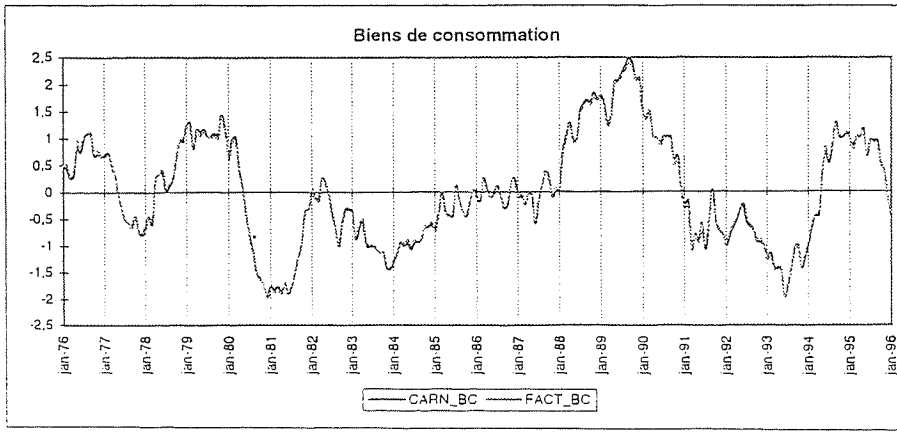




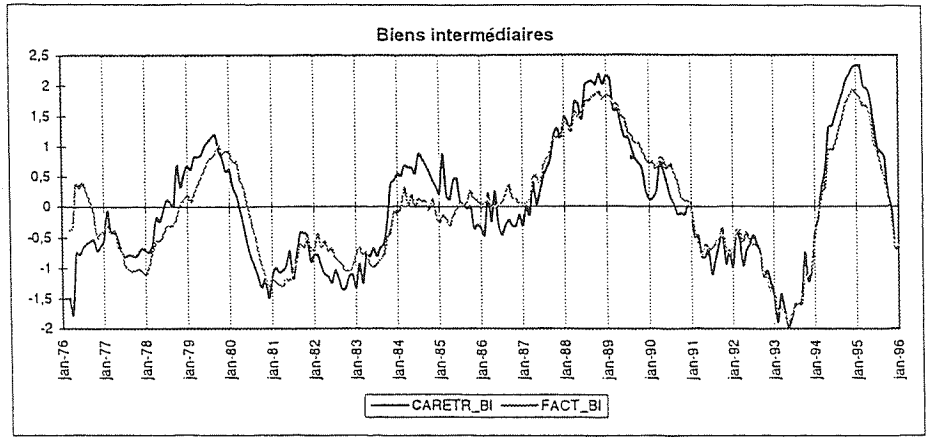
GRAPHIQUE 8.3 : CARNETS DE COMMANDE
Données initiales et facteurs communs par branche

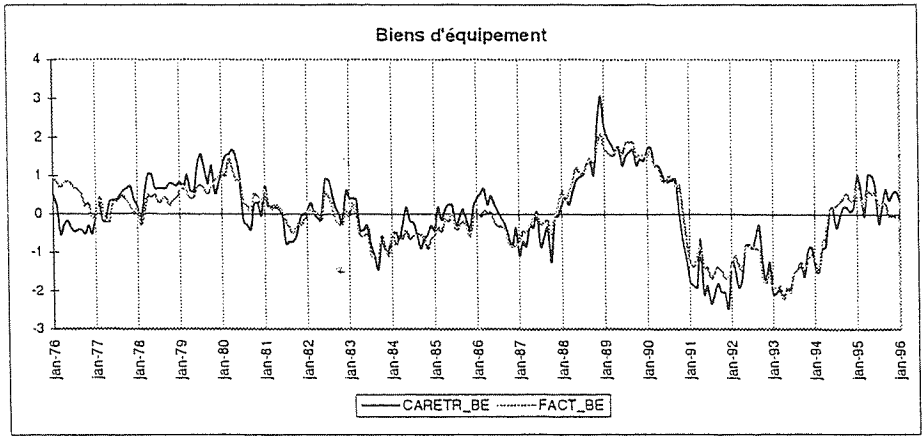


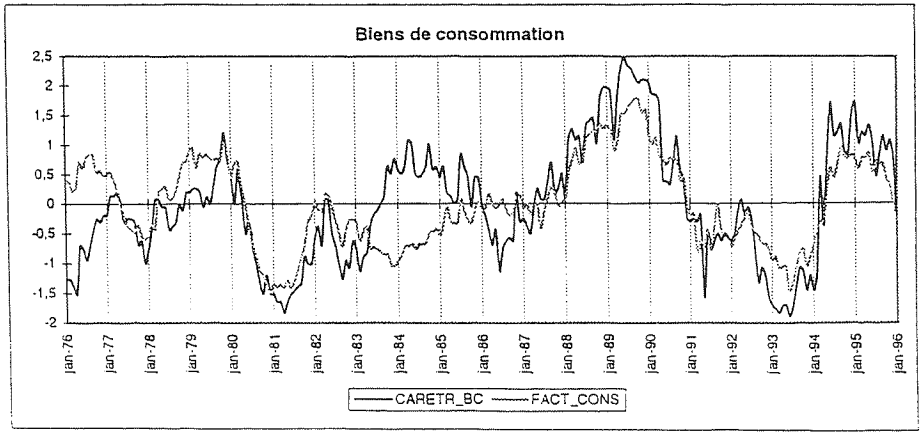




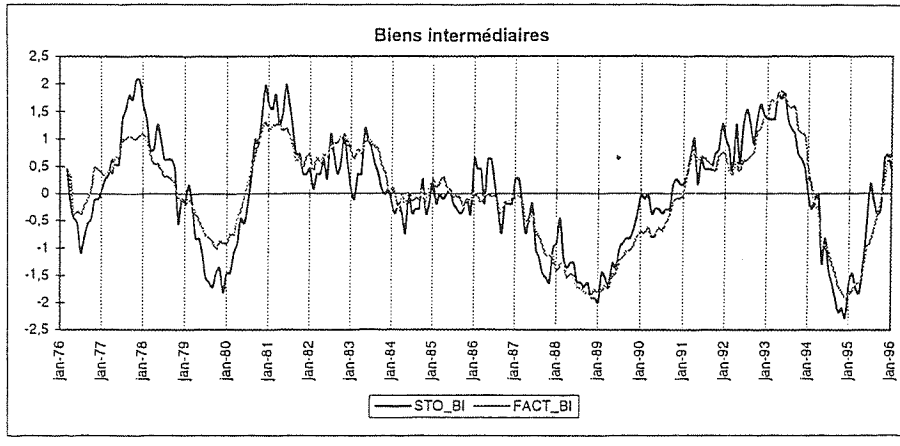
GRAPHIQUE 8.4 : CARNETS DE COMMANDE ETRANGERS
Données initiales et facteurs communs par branche

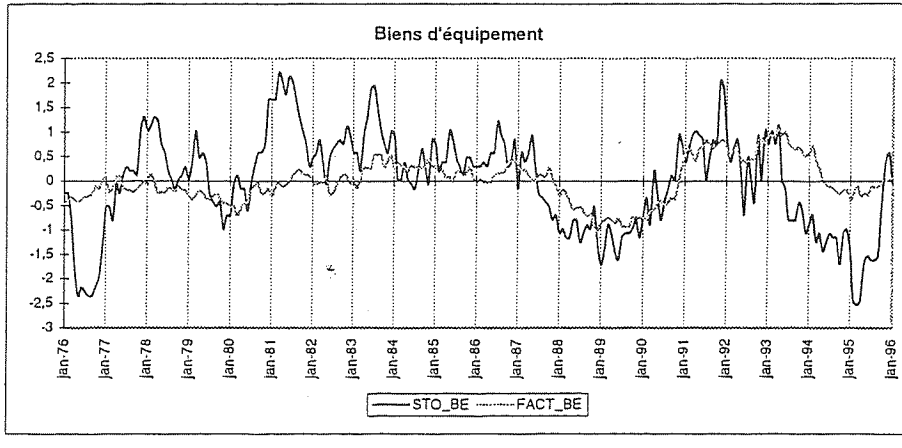


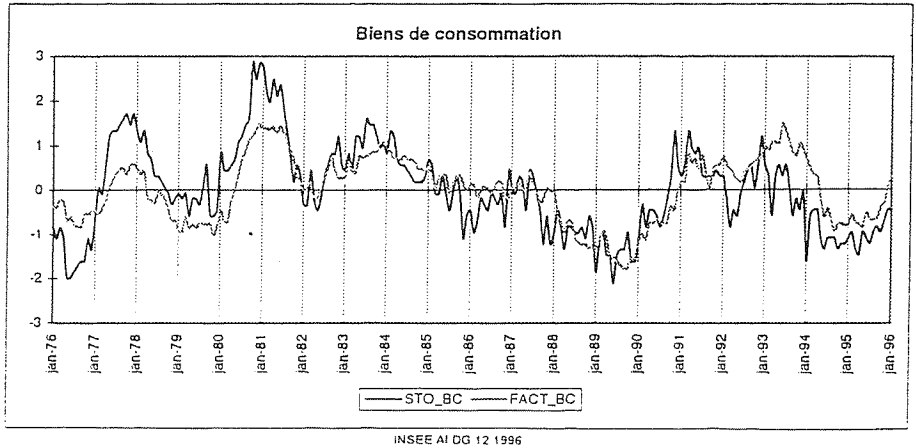




GRAPHIQUE 8.5 : OPINION SUR LES STOCKS
Données initiales et facteurs communs par branches







INSEE AI DG 12 1996

Session 3

Les mesures d'inégalité

LES PRINCIPALES MESURES D'INÉGALITÉ

Olivier Sautory

Avertissement

Ce papier constitue une version provisoire d'un document plus ambitieux, qui contiendra en particulier d'autres mesures d'inégalité (peut-être un peu plus "exotiques" que celles abordées dans ce papier...), des exemples d'utilisation des mesures présentées ici, les macros SAS mises à la disposition des statisticiens de l'INSEE pour les calculer, une bibliographie, une intoduction (et une conclusion) digne de ce nom, etc.

L'auteur tient à préciser que deux documents ont très largement constitué la "matière première" qui a permis d'élaborer ce papier, à savoir :

Gouriéroux C., 1981, "Mesures d'inégalité, de pauvreté, de concentration", cours polycopié, ENSAE.

Villeneuve A., 1984, "La mesure de l'inégalité par les indicateurs de concentration (présentation des indicateurs les plus connus et de divers problèmes et propriétés)", Note interne INSEE n°1380/450 du 7 juin 1984.

La présentation des mesures d'inégalité adoptée ici relève beaucoup plus de la "statistique descriptive" que de considérations de nature économique, voire philosophique, qui leur sont associées. La (future) bibliographie renverra les lecteurs intéressés par ces aspects aux textes qui les abordent explicitement (les deux documents mentionnés plus haut contiennent d'ailleurs certains développements sur ces sujets).

I. Courbe de Lorenz

On considère une population de n individus supposés numérotés par revenu croissant :

$$x_1 \leq x_2 \leq \ldots \leq x_i \leq \ldots \leq x_n$$

où x_i désigne le revenu de l'individu i.

On note :

- $m = \frac{1}{n} \sum_{i=1}^{n} x_i$ le revenu moyen dans la population

- $y_i = \frac{x_i}{m}$ le revenu "relatif" (i.e. rapporté à la moyenne) de l' individu i.

I.1 Définition

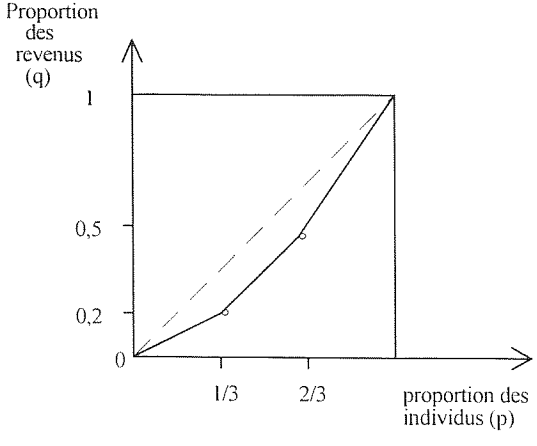
On note q_k la proportion du revenu total reçue par les k individus les plus pauvres :

$$q_k = \frac{x_1 + \ldots + x_k}{x_1 + \ldots + x_n} = \frac{1}{n} \sum_{i=1}^{k} y_i \quad (1 \leq k \leq n)$$

(on pose $q_0 = 0$)

La courbe de Lorenz est la courbe reliant les points $\left(p_k = \frac{k}{n}, q_k \right)$, k = 0, 1 ...n

Exemple : n = 3 $x_1 = 2$ $x_2 = 3$ $x_3 = 5$



$p_1 = \frac{1}{3}$ $p_2 = \frac{2}{3}$ $p_3 = 1$

$q_1 = 0{,}2$ $q_2 = 0{,}5$ $q_3 = 1$

I.2 Propriétés

I.2.1 Forme de la courbe - Invariance

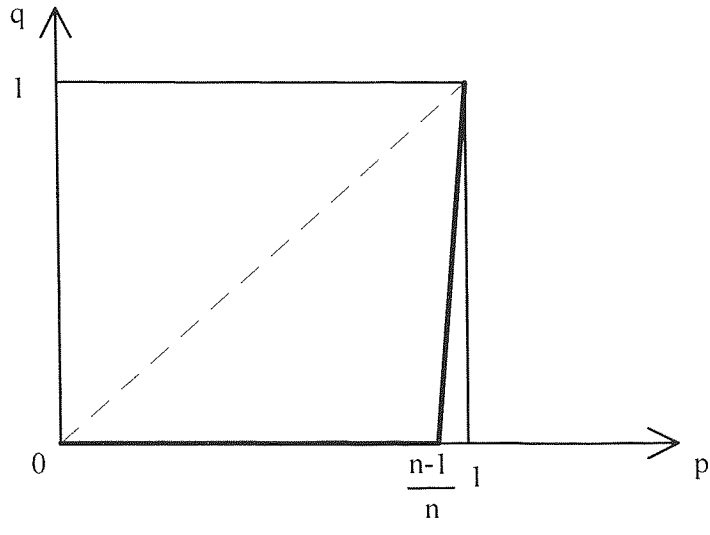
La courbe de Lorenz passe par les points (0,0) et (1,1), elle est linéaire par morceaux, inscrite dans le carré de côté 1, croissante (strictement si $x_1 > 0$, convexe (car $q_{k+1} - q_k \geq q_k - q_{k-1}$) donc située sous la première bissectrice.

La courbe de Lorenz est invariante par changement d'échelle (ou d'unité) sur les revenus. En revanche si tous les revenus sont augmentés d'une quantité positive, la courbe de Lorenz de la nouvelle distribution est située au-dessus de l'ancienne.

I.2.2 Cas limite

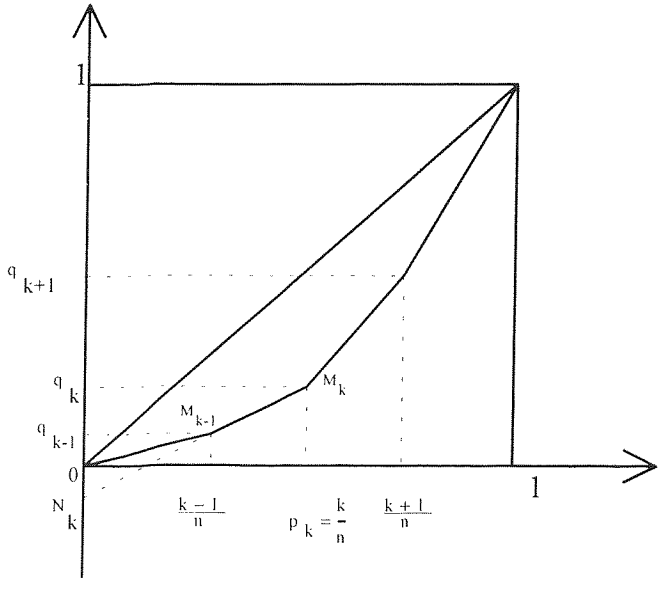
Dans une distribution égalitaire, tous les individus reçoivent le même revenu, et les p% les plus "pauvres" reçoivent p% du revenu total : la courbe de Lorenz est confondue avec la bissectrice. Une courbe proche de cette bissectrice correspond à une distribution presque égalitaire ; inversement, une distribution fortement inégalitaire est caractérisée par une courbe de Lorenz proche des côtés du carré.

L'inégalité est maximale si un seul individu a un revenu non nul, tous les autres ayant un revenu nul. La courbe de Lorenz correspondante à la forme suivante :



I.2.3 Propriétés géométriques

Dans tout ce paragraphe, on raisonne en termes de revenus **relatifs** y_i .



On note M_k le point courant (p_k, q_k) de la courbe de Lorenz.

• La pente de la droite OM_k vaut :

$$\frac{q_k}{p_k} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

C'est le revenu moyen des k individus les plus pauvres.

• La pente de la droite $M_{k-1}M_k$ vaut :

$$\frac{q_k - q_{k-1}}{p_k - p_{k-1}} = y_k$$

C'est le revenu du k-ième individu.

• Soit N_k le point d'intersection de la droite $M_{k-1}M_k$, d'équation $Y - q_k = y_k (X - p_k)$, avec l'axe des ordonnées ; la longueur du segment ON_k vaut :

$$-q_k + p_k y_k = -\frac{1}{n}\sum_{i=1}^{k} y_i + \frac{k}{n} y_k = \frac{1}{n}\sum_{i=1}^{k} (y_k - y_i)$$

C'est, au coefficient $\frac{1}{n}$ près, la quantité de revenus à verser aux k - 1 individus les plus pauvres pour leur assurer un revenu égal à celui du k-ième individu.

I.3 Données regroupées en classes

Il arrive fréquemment que l'on ne dispose que des données regroupées en H classes (et non pas des données individuelles) :

$$[e_0 = 0, e_1[\ldots [e_{h-1}, e_h[\ldots [e_{H-1}, e_H = +\infty[$$

Soit n_h le nombre d'individus ayant un revenu compris entre e_{h-1} et e_h, et X_h la somme des revenus de ces n_h individus. On construit alors la courbe de Lorenz pour les valeurs correspondant aux extrémités des classes :

$$p(e_h) = \frac{n_1 + \ldots + n_h}{n} \qquad q(e_h) = \frac{X_1 + \ldots + X_h}{n\,m} \qquad (h = 1 \ldots H)$$

La fraction $p(e_h)$ de la population reçoit la fraction $q(e_h)$ du revenu total nm.

Pour tracer la courbe, on relie ces points entre eux par des segments de droite. La courbe de Lorenz a donc la même forme que celle obtenue au § I.1, et les propriétés sont identiques à celles énoncées au § I.2. Les propriétés vues au § I.2.3 deviennent ici, en notant M_h le point courant $(p(e_h), q(e_h))$ de la courbe :

• la pente de la droite OM_h est le revenu (relatif) moyen des $n\,p(e_h)$ individus les plus pauvres ;

• la pente de la droite $M_{h-1}M_h$ est le revenu moyen des individus ayant un revenu compris entre e_{h-1} et e_h ;

• la longueur du segment ON_h est, au coefficient $\frac{1}{n}$ près, la quantité de revenus à verser aux $n\ p(e_{h-1})$ individus les plus pauvres pour leur assurer un revenu égal au revenu moyen des individus de la h-ième classe.

I.4 Généralisation à une variable continue

On peut construire la courbe de Lorenz pour toute variable numérique continue positive, de densité f (> 0) et de fonction de répartition F (inversible).

I.4.1 Définition

Une proportion d'individus p(x) = F(x) (avec x > 0) reçoit une proportion du revenu total égale à :

$$q(x) = \frac{\int_0^x u\, f(u)\, du}{\int_0^\infty u\, f(u)\, du} = \frac{1}{m}\int_0^x u\, f(u)\, du$$

où m est le revenu moyen dans la population.

La courbe de Lorenz est la courbe paramétrée : $(p = p(x),\ q = q(x))\quad (x > 0)$, soit

$$q = L(p) = \frac{1}{m}\int_0^{F^{-1}(p)} u\, f(u)\, du = \frac{1}{m}\int_0^p F^{-1}(u)\, du$$

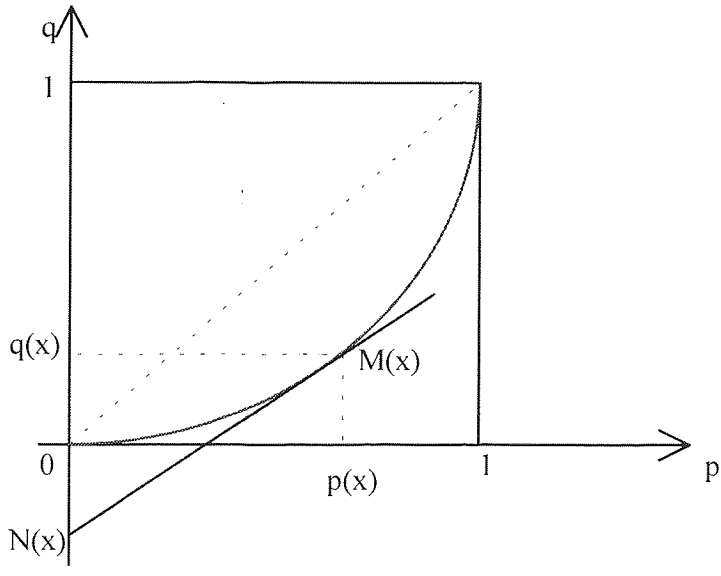
I.4.2 Propriétés

On peut retrouver à l'aide de l'expression précédente les propriétés vues au § I.2 :

• $L(0) = 0,\ L(1) = 1$

• Croissance : $L'(p) = \frac{1}{m} F^{-1}(p) > 0$

• Convexité : $L''(p) = \frac{1}{f\left[F^{-1}(p)\right]} > 0$



• pente de $OM(x) = \dfrac{q(x)}{F(x)} =$ revenu (relatif) moyen des individus ayant un revenu inférieur ou égal à x.

• pente de la tangente à la courbe en $M(x)$:

$$\frac{1}{m} F^{-1}(p) = \frac{x}{m} = \text{revenu (relatif) en ce point.}$$

• longueur du segment $ON(x) = x\ p(x) - q(x) =$ quantité du revenu à accorder aux plus pauvres (de revenu < x) pour leur assurer un revenu égal à x.

I.4.3 Exemple

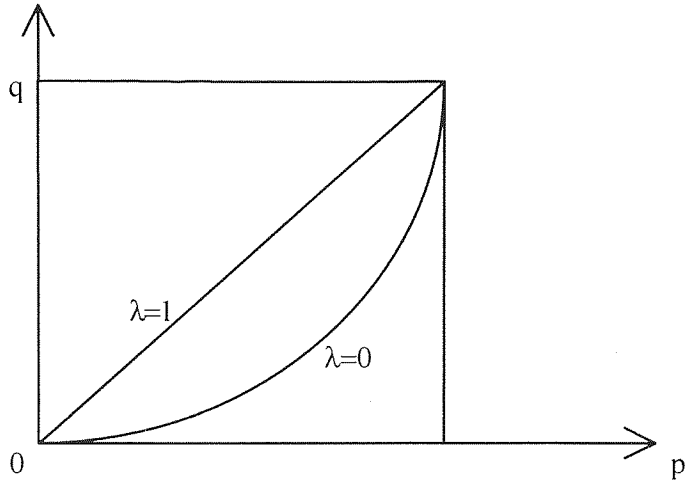
On suppose que la variable revenu suit une loi uniforme sur [a, b] (0 < a < b).

On a alors, pour x ∈ [a, b] :

$$p(x) = \frac{x-a}{b-a}, \quad q(x) = \frac{x^2 - a^2}{b^2 - a^2}$$

d'où :

$$q = \frac{(1-\lambda)p^2 + 2\lambda p}{1+\lambda}, \text{ avec } \lambda = \frac{a}{b} \in]0, 1[$$



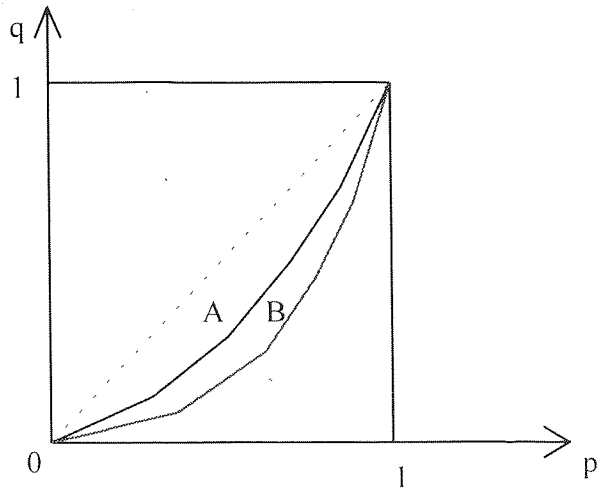
Le paramètre λ s'interprète comme une mesure d'inégalité : plus il est élevé, plus la distribution est égalitaire.

Le cas $\lambda = 1$ $(i.e.\, b \to a)$ correspond à la première bissectrice : tous les revenus sont égaux ; le cas $\lambda = 0$ $(i.e.\, b \to +\infty)$ correspond à la partie de la parabole $q = p^2$: l'inégalité est maximale.

Remarque : relier les points successifs par des segments de droite dans le cas de données regroupées en classes ne revient donc pas à supposer une répartition uniforme des revenus dans chaque classe.

I.5 Comparaison des courbes de Lorenz

Soient deux distributions de revenus A et B, pas nécessairement relatives à une même population, telles que la courbe de Lorenz soit toujours en-dessous de celle de A (avec éventuellement des points communs).



Cela signifie que les p% individus les plus pauvres de A sont, quel que soit p, moins pauvres que ceux de B (i.e. leur revenu moyen est plus élevé dans la distribution A que dans la distribution B). On dira dans ce cas que **la distribution B est plus inégalitaire que la distribution A**.

Si on considère maintenant que ces distributions sont relatives à une même population, et correspondent à deux partages d'un même revenu total (ce que l'on peut supposer, car une courbe de Lorenz ne dépend pas du nombre n de revenus, et est invariante par homothétie), on a la propriété suivante : la distribution B est plus inégalitaire que A si et seulement si A peut être obtenue à partie de B par une suite de "transferts" de riches vers les pauvres, un transfert étant défini de la façon suivante :

un transfert égal à h (> 0) du riche j vers le pauvre i (< j) est tel que le revenu de j devient $x_j - h$ et le revenu de i devient $x_i + h$ (avec $h < x_j - x_i$), les autres revenus restant inchangés.

II. Mesures scalaires d'inégalités

La relation d'inégalité fondée sur la comparaison des courbes de Lorenz est une relation d'ordre partiel sur les distributions : on ne peut comparer deux distributions pour lesquelles les courbes de Lorenz correspondantes ont un point d'intersection. Il

est donc nécessaire, pour comparer deux distributions quelconques, de définir une mesure scalaire d'inégalité I :

$$distribution\ x\ =\ (x_1, \ldots, x_n) \rightarrow I(x) \in IR$$

On dira qu'une distribution x est plus inégalitaire (au sens de la mesure I) qu'une distribution z si I(x)>I(z).

II.1 Condition de Pigou-Dalton (Lorenz)

On cherchera (généralement) des mesures d'inégalité I telles que le préordre total associé à I soit compatible avec le préordre partiel déduit de la comparaison des courbes de Lorenz, i.e. des mesures I vérifiant les conditions suivantes (qui sont équivalentes) :

- "de Lorenz" : si x est plus inégalitaire que z (au sens défini au § 1.5), alors I (x) > I (z);

- "de Pigou-Dalton" : un transfert d'un riche vers un pauvre entraîne une diminution de I (si l'écart entre les deux revenus considérés a diminué).

Remarque : on peut définir une condition de Pigou-Dalton-Lorenz au sens large :

- x plus inégalitaire que z $\Rightarrow$ I (x) $\geq$ I (z) ;

- un transfert d'un riche vers un pauvre ne peut s'accompagner d'une augmentation de I.

Transfert

Lorsque l'on parlera de transfert par la suite, il s'agira toujours de revenus **relatifs** :

$$y_j \rightarrow y_j^* = y_j - h$$
$$y_i \rightarrow y_i^* = y_i + h$$

avec $y_i < y_j$ *et* $h < y_j - y_i$.

On peut noter qu'à l'issue du transfert l'individu j peut se retrouver plus pauvre que l'individu i ($si\ h > \frac{1}{2}(y_j - y_i)$), mais l'écart entre leurs revenus est moins grand : $\left|y_j^* - y_i^*\right| < y_j - y_i$.

Pour étudier les conséquences d'un transfert sur les différentes mesures d'inégalité que l'on va présenter, il sera parfois commode de supposer le transfert "infinitésimal". La variation d'une mesure I s'écrira alors :

$$dI = I\left(y_1..(y_i+h)..(y_j-h)..y_n\right) - I\left(y_1..y_i..y_j..y_n\right) = h\left(\frac{\partial I}{\partial y_i}\left(y_1..y_n\right) - \frac{\partial I}{\partial y_j}\left(y_1..y_n\right)\right)$$

II.2 Condition suffisante

La condition de Pigou-Dalton-Lorenz est vérifiée si I est de la forme :

$$I = -\sum_{k=1}^{n} u\left(\frac{x_k}{n}\right) = -\sum_{k=1}^{n} u(y_k)$$

où u est une fonction strictement concave.

On a alors en effet : $dI = -h\left(u'(y_i) - u'(y_j)\right) < 0$

puisque u' est décroissante.

Interprétation économique

- u s'interprète comme une fonction d'utilité individuelle, dépendant uniquement du revenu, la même pour tous les individus.

- $W(x_1 \ldots x_n) = \sum_{k=1}^{n} u(y_k)$ s'interprète comme une fonction d'utilité collective, dépendant uniquement des niveaux d'utilité de chaque individu, ayant comme propriétés d'être :

  - croissante dans les utilités individuelles ;
  - symétrique (exigence d'anonymat) ;
  - quasi-concave (aversion pour l'inégalité).

I = - W peut dont s'interpréter comme une mesure d'inégalité.

III. Indice de Gini

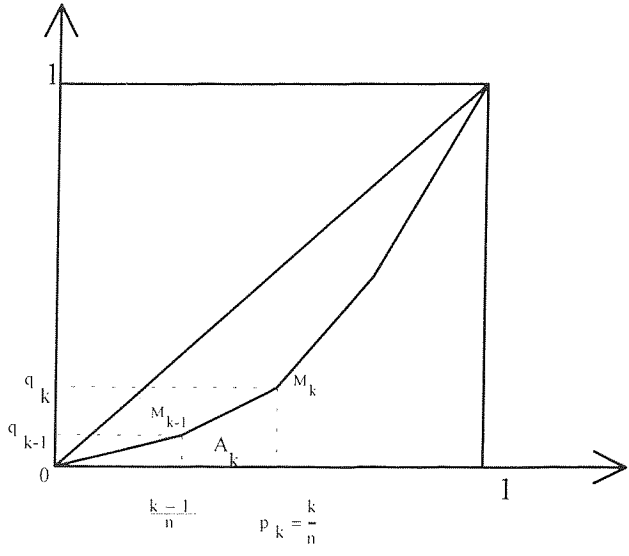
III.1 Définition

L'indice de Gini est égal au double de l'aire comprise entre la courbe de Lorenz et la bissectrice.

Cette mesure d'inégalité, la plus utilisée, a été introduite par le statisticien italien Corrado Gini.

III.2 Expressions de l'indice de Gini

III.2.1. Données individuelles



L'aire A_k vaut :

$$A_k = (q_k - q_{k-1}) \times \frac{1}{2}\left[\left(1 - \frac{k}{n}\right) + \left(1 - \frac{k-1}{n}\right)\right] = \frac{1}{2n^2} y_k (2n - 2k + 1)$$

On en déduit la valeur de l'indice de Gini :

$$G = 2\times\left(\frac{1}{2} - \sum_{k=1}^{n} A_k\right) = 1 - \frac{1}{n^2}\sum_{k=1}^{n} y_k\,(2n - 2k + 1) \qquad (1)$$

$$= \frac{1}{n}\sum_{k=1}^{n} y_k - \frac{1}{n^2}\sum_{k=1}^{n} y_k\,(2n - 2k + 1)$$

$$\boxed{G = \frac{1}{n^2}\sum_{k=1}^{n} y_k\,(2k - n - 1)} \qquad (2)$$

En utilisant la relation :

$$\sum_{k}\sum_{k'} min\left(y_k, y_{k'}\right) = \sum_{k} y_k + 2\sum_{k}\sum_{k'>k} min\left(y_k, y_{k'}\right) = \sum_{k} y_k + 2\sum_{k}(n-k)\,y_k$$

on peut déduire de (1) :

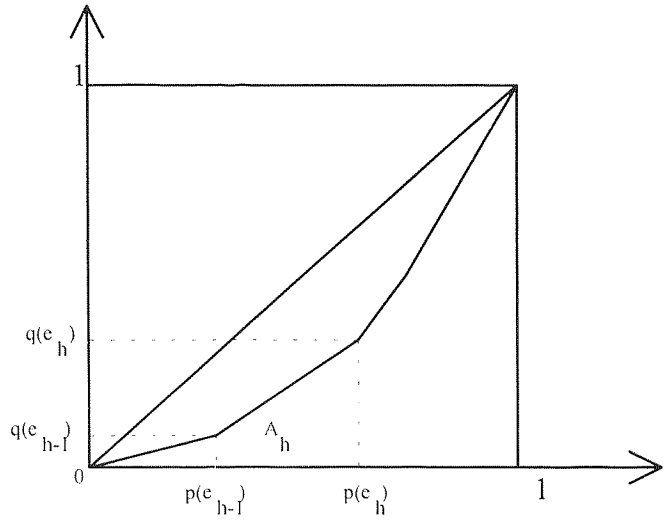
$$\boxed{G = 1 - \frac{1}{n^2}\sum_{k=1}^{n}\sum_{k'=1}^{n} min\left(y_k, y_{k'}\right)} \qquad (3)$$

Si on utilise la relation $y_k + y_{k'} = 2\,min\,(y_k, y_{k'}) + \left|y_k - y_{k'}\right|$, on obtient aussi :

$$\boxed{G = \frac{1}{2n^2}\sum_{k=1}^{n}\sum_{k'=1}^{n}\left|y_k - y_{k'}\right|} \qquad (4)$$

III.2.2. Données regroupées en classes

On conserve les mêmes notations que dans le § I.3.



L'aire A_h vaut : $\left[q(e_h)-q(e_{h-1})\right]\left[1-\frac{p(e_{h-1})+p(e_h)}{2}\right]$

On en déduit la valeur de $G = 2\left(\frac{1}{2}-\sum_h A_h\right)$:

$$\boxed{G=\sum_{h=1}^{H}\left[q(e_h)-q(e_{h-1})\right]\left[p(e_h)+p(e_{h-1})\right]-1}$$

On obtient aussi, de façon analogue :

$$\boxed{G=1-\sum_{h=1}^{H}\left[p(e_h)-p(e_{h-1})\right]\left[q(e_h)+q(e_{h-1})\right]}$$

Comme on l'a vu au § I.4.3, l'indice G ainsi calculé est un minorant de l'indice G_0 que l'on obtiendrait si l'on connaissait les revenus individuels. On peut obtenir un majorant de G_0 en utilisant la propriété de convexité de la courbe de Lorenz.

III.2.3. Cas d'une variable continue

On conserve les mêmes notations que dans le § I.4.

L'aire comprise entre la première bissectrice et la courbe de Lorenz q = L (p) vaut :

$$\int_0^1 \left[p-L(p)\right]dp.$$

On en déduit la valeur de l'indice de Gini :

$$\boxed{G = 2\int_0^1 [p - L(p)dp] = 1 - 2\int_0^1 L(p)dp}$$

On peut aussi écrire :

$$G = 1 - \frac{2}{m}\int_0^1 \left(\int_0^p F^{-1}(u)\,du\right)dp = 1 - \frac{2}{m}\int_0^1 (1-u)\,F^{-1}(u)\,du$$

soit, en utilisant la relation $\int_0^1 F^{-1}(u)\,du = m$:

$$\boxed{G = -1 + \frac{2}{m}\int_0^1 u\,F^{-1}(u)\,du = -1 + \frac{2}{m}\int_0^\infty v\,F(v)\,f(v)\,dv}$$

On peut aussi obtenir une formule équivalente à l'expression (4) du § III.2.1 :

$$\boxed{G = \frac{1}{2m}\int_0^\infty \int_0^\infty |u - v|\,f(u)\,f(v)\,du\,dv}$$

Exemple

Si la variable revenu suit une loi uniforme sur [a, b], 0 < a < b :

$$G = 1 - 2\int_0^1 \frac{(1-\lambda)p^2 + 2\lambda p}{1+\lambda}dp = \frac{1}{3}\frac{1-\lambda}{1+\lambda} \quad avec\ \lambda = \frac{a}{b}$$

III.3 Valeurs minimale et maximale

- G est minimal, et vaut alors 0, si la courbe de Lorenz coïncide avec la première bissectrice, i.e. dans le cas d'une distribution parfaitement égalitaire.

- G est maximal si tous les y_i sont nuls sauf un : $y_1 = \ldots = y_{n-1} = 0 \qquad y_n = n$

$$G_{max} = \frac{n-1}{n}$$

III.4 Interprétations

1. Par sa définition même, G s'interprète comme une certaine "distance" à la situation d'égalité.

2. D'après la formule (4) du § III.2.1, G est la demi-moyenne des écarts absolus entre les y_i pris 2 à 2 : c'est donc un indicateur de dispersion.

3. On peut réécrire la formule (2) du § III.2.1 sous la forme suivante :

$$G = \frac{2}{n} \sum_{k=1}^{n} (y_k - 1)\left(\frac{k}{n} - \frac{n+1}{2n}\right)$$

G/2 est égal à la covariance entre le niveau relatif des revenus (variable y_k) et le rang relatif dans l'échelle des revenus (variable k/n).

4. L'indice de Gini peut également s'interpréter en termes de perceptions individuelles du phénomène de l'inégalité (cf. § II).

   - La formule (4) du § III.2.1 s'écrit :

$$G = -\frac{1}{n} \sum_{k=1}^{n} u_1(y_k) \qquad avec\ u_1(y_k) = -\frac{1}{2n} \sum_{k'=1}^{n} \left| y_k - y_{k'} \right|$$

L'utilité de l'individu k est mesurée par les écarts entre son revenu et ceux des autres individus.

   - Cette formule peut se réécrire :

$$G = \frac{1}{n^2} \sum_{k=1}^{n} \left(\sum_{k'<k} (y_k - y_{k'}) \right) = \frac{1}{n^2} \sum_{k=1}^{n} \left((k-1) y_k - \sum_{k'<k} y_{k'} \right)$$

$$= \frac{1}{n} \sum_{k=1}^{n} \frac{k-1}{n} \left(y_k - \frac{1}{k-1} \sum_{k'<k} y_{k'} \right) = -\frac{1}{n} \sum_{k=1}^{n} u_2(y_k)$$

$$\text{avec } u_2(y_k) = -\frac{k-1}{n} \left(y_k - \bar{y}_{inf}(k) \right),$$

où $\bar{y}_{inf}(k)$ est la moyenne des revenus inférieurs à y_k.

L'utilité de l'individu k est le produit de la proportion d'individus plus pauvres que lui par l'écart entre son propre revenu et leur revenu moyen.

- On peut aussi écrire, de façon analogue :

$$G = \frac{1}{n} \sum_{k=1}^{n} \frac{n-k}{n} \left(\frac{1}{n-k} \sum_{k'>k} \left(y_{k'} - y_k \right) \right) = -\frac{1}{n} \sum_{k=1}^{n} u_3(y_k)$$

avec $u_3(y_k) = -\frac{n-k}{n}(\bar{y}_{sup}(k) - y_k)$,

où $\bar{y}_{sup}(k)$ est la moyenne des revenus supérieurs à y_k

L'utilité de l'individu k est le produit de la proportion d'individus plus riches que lui par l'écart entre leur revenu moyen et son propre revenu.

III.5 Conséquence d'un transfert

En utilisant la formule (2) du § III.2.1, il est facile de vérifier que la variation de l'indice de Gini consécutive à un transfert vaut, en supposant h assez petit pour ne pas affecter les rangs du " riche" et du "pauvre" :

$$\Delta G = -\frac{2}{n^2}(j-i)h$$

Comme attendu d'après sa définition même, l'indice de Gini vérifie la condition de Pigou-Dalton-Lorenz ; la variation ΔG dépend des rangs des individus, et non de leurs revenus. Cet indice accorde donc le même poids à l'inégalité parmi les riches ou parmi les pauvres.

IV. Autres mesures d'inégalité

IV.1 Écart relatif moyen

IV.1.1 Définition

Cette mesure est la moyenne arithmétique des valeurs absolues des écarts des revenus (relatifs) à la moyenne :

$$D = \frac{1}{n\bar{x}} \sum_{k=1}^{n} |x_k - \bar{x}| = \frac{1}{n} \sum_{k=1}^{n} |y_k - 1|$$

Dans le cas d'une variable continue, D a pour expression :

$$D = \frac{1}{n} \int_0^{\infty} |u - m| \, f(u) \, du$$

Remarque : on rencontre dans la littérature **l'indicateur de Kuznets**, noté K, qui n'est en fait rien d'autre que D/2.

Valeurs minimale et maximale

- $D_{min} = 0$: tous les revenus sont égaux
- $D_{max} = \frac{n-1}{n}$: tous les revenus sont nuls, sauf un.

IV.1.2 Interprétations

1. D est le *pourcentage d'égalisation maximale*, i.e. la proportion (minimale) de la masse totale des revenus $(n\bar{x})$ qu'il faudrait redistribuer entre les individus pour atteindre l'égalité complète (par les transferts $|x_k - \bar{x}|$).

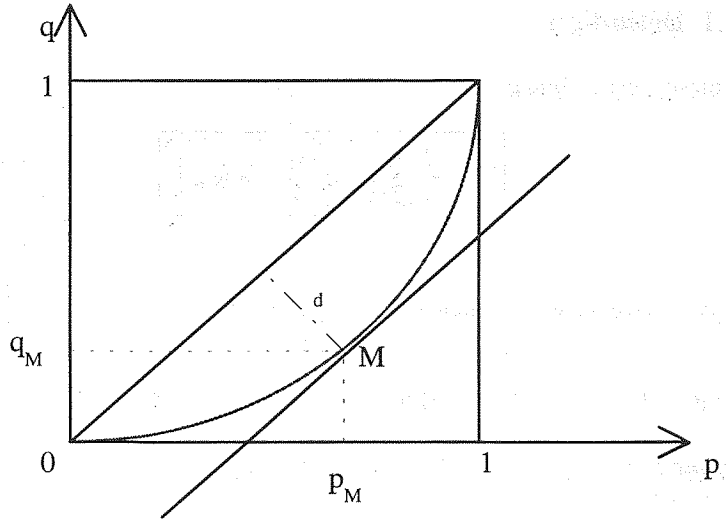
2. K (=D/2) est égal à la différence entre la proportion de la masse totale des revenus détenue par les individus "riches" (i.e. de revenu supérieur à la moyenne) et la proportion de ces individus "riches" dans la population. K est aussi égal à la différence entre la proportion d'individus "pauvres" (i.e. de revenu inférieur à la moyenne) et la proportion de la masse totale des revenus détenue par ces individus.

$$K = \frac{n_r \bar{x}_r}{n\bar{x}} - \frac{n_r}{n} = \frac{n_p}{n} - \frac{n_p \bar{x}_p}{n\bar{x}}$$

où n_r et n_p désignent les nombres de riches et de pauvres, $\bar{x}_r$ *et* $\bar{x}_p$ les revenus moyens des riches et des pauvres (voir démonstration en annexe).

3. D est la "distance" (à une constante multiplicative près) entre la courbe de Lorenz de la distribution et la première bissectrice.

$$D = 2\sqrt{2}d$$



M est le point de la courbe de Lorenz où la tangente est parallèle à la bissectrice : c'est le point où la distance à cette bissectrice est maximale (en raison de la convexité de la courbe).
(voir démonstration en annexe).

IV.1.3 Conséquence d'un transfert

D'après la définition de D, et en utilisant l'interprétation 2 du § IV.1.2, il est facile de vérifier que la variation de l'écart relatif consécutive à un transfert vaut :

$\Delta D = -\frac{2h}{n}$ *si* $y_i < l < y_j$ i.e. si l'individu j a un revenu supérieur à la moyenne et l'individu i a un revenu inférieur à la moyenne

$= 0$ sinon, i.e. si les individus i et j sont tous les deux "riches" ou tous les deux "pauvres".

La condition de Pigou-Dalton-Lorenz n'est donc vérifiée qu'au sens large.

L'interprétation 3 du § IV.1.2 permet d'illustrer le fait que la condition de Lorenz n'est vérifiée qu'au sens large : deux distributions peuvent avoir des courbes de Lorenz ayant comme point commun le point M défini ci-dessus, l'une étant néanmoins toujours située sous l'autre. L'une des deux distributions est donc plus inégalitaire que l'autre, bien qu'elles conduisent à la même valeur de D.

IV.2 Variance des logarithmes

IV.2.1 Définition

Cette mesure est définie par :

$$VL = \frac{1}{n} \sum_{k=1}^{n} \left(Log \frac{x_k}{\bar{x}} - Log\, g \right)^2$$

où g est la moyenne géométrique des revenus relatifs $y_k \left(= x_k / \bar{x}\right)$: $g = \left(\prod_{k=1}^{n} y_k \right)^{\frac{1}{n}}$

VL est donc la variance des logarithmes des revenus relatifs.

Remarque : on trouve parfois dans la littérature une formule erronée de la variance des logarithmes :

$$VL' = \frac{1}{n} \sum_{k=1}^{n} \left(Log \frac{x_k}{\bar{x}} \right)^2$$

Dans cette formule, la moyenne des logarithmes des revenus (Log g) est remplacée par le logarithme du revenu moyen (qui vaut 0).

Valeurs minimales et maximales

- $VL_{min} = 0$: tous les revenus sont égaux.
- $VL \rightarrow + \infty$ dès qu'un revenu x_k est nul.

Cette propriété rend l'indicateur inutilisable lorsqu'il y a des revenus nuls (en particulier dans le cas où ces revenus nuls sont assez fréquents dans la population, par exemple, pour les revenus fonciers, immobiliers ...). Même si le calcul ne se fait

pas au niveau individuel (mais décile par décile par exemple), l'indicateur VL sera sensible aux très petites valeurs (voir aussi plus loin), qui ne sont d'ailleurs pas en général les plus fiables.

IV.2.2 Conséquence d'un transfert

La variation de VL consécutive à un transfert infinitésimal h d'un riche j vers un pauvre i vaut :

$$d\,VL = -\frac{2h}{n}\left(\frac{Log\left(y_j/g\right)}{y_j} - \frac{Log\left(y_i/g\right)}{y_i}\right)$$

On a donc :

$d\,VL < 0$ si $y_i < y_j \leq eg$

$d\,VL > 0$ si $y_j > y_i \geq eg$

où $e \approx 2{,}7$

(voir démonstration en annexe).

VL ne vérifie donc pas la condition de Pigou-Dalton-Lorenz ; le résultat est même aberrant si le transfert s'effectue entre deux individus dont les revenus sont supérieurs à 2,7 fois le revenu moyen (car $g \approx 1$).

Remarques

- Il est facile de vérifier que $f''(y) < 0 \; si \; y < eg$: à différence de revenus $y_j - y_i$ constante $\left(avec\; y_i < y_j < eg\right)$, la diminution de VL est d'autant plus forte que les revenus sont faibles. Cet indicateur accorde donc un poids plus important à l'inégalité parmi les pauvres.

- La condition de Pigou-Dalton-Lorenz n'est pas vérifiée non plus si l'on utilise la "fausse" variance des logarithmes VL' (remplacer g par 1 dans les formules précédentes).

Conclusion

La variance des logarithmes n'est pas un bon indicateur. Il est néanmoins couramment utilisé, pour des raisons :

• ***historiques*** : son emploi a été proposé il y a longtemps, avec l'idée que s'intéresser aux logarithmes des revenus, et à leur dispersion mesurée par la variance, c'est étudier les différences **relatives** de revenus ;

• ***théoriques*** : il est "classique" d'admettre qu'une distribution de revenus suive un loi log-normale, pour laquelle la variance des logarithmes est un indicateur de dispersion tout à fait adéquat.

IV.3 Indicateurs d'Atkinson

IV.3.1 Définition

L'indicateur d'Atkinson A(a), avec a < 1, est défini par :

$$A(a) = 1 - \left(\frac{1}{n} \sum_{k=1}^{n} (y_k)^a \right)^{\frac{1}{a}} = 1 - m_a$$

où m_a est la *moyenne d'ordre a* de la distribution des y_k.

Cas particuliers

• a = - 1

A (-1) = 1 - H, où H est la moyenne harmonique des y_k.

• a = 0

La formule donnée est indéterminée, mais par continuité m_a tend vers la moyenne géométrique G des y_k ; on posera donc :

$$A(0) = 1 - G$$

Cet indicateur est aussi appelé **indicateur de Champernowne**.

On a : $- Log\,[1 - A(0)] = -Log\,G = -\frac{1}{n}\sum_{k=1}^{n} Log\,y_k$.

On rencontre cet indicateur dans la littérature sous le nom de déviation logarithmique moyenne, ou **écart moyen des logarithmes**.

Valeurs minimale et maximale

- $A(a)_{min} = 0$: tous les revenus sont égaux.

$$(car\ \frac{1}{n}\sum_{k=1}^{n}(y_k)^a < \frac{1}{n}\sum_{k=1}^{n} y_k = 1, \text{ sauf si } y_k = 1\ \forall k)$$

- Pour la valeur maximale de A(a), il faut distinguer 2 cas :

1er cas : $a \leq 0$

$A(a) \rightarrow 1$ dès qu'un revenu x_k est nul.

Ce résultat disqualifie donc cet indicateur pour certains types de revenus, comme c'était le cas pour la variance des logarithmes.

2ème cas : $a > 0$

La valeur maximale de A(a), obtenue lorsque tous les revenus sont nuls sauf un, vaut :

$$A(a)_{max} = 1 - n^{\frac{a-1}{a}}$$

(voir démonstration en annexe)

IV.3.2 Conséquence d'un transfert

La variation de A(a) consécutive à un transfert infinitésimal h d'un riche j vers un pauvre i vaut :

$$dA(a) = \frac{(1-A)^{1-a}h}{n}\left(\left(y_j\right)^{a-1} - \left(y_i\right)^{a-1}\right) < 0$$

(voir démonstration en annexe).

Cette formule est valable aussi lorsque a=0. La condition de Pigou-Dalton-Lorenz est vérifiée pour tout indicateur d'Atkinson A(a) (a < 1), puisque $(y_j)^{a-1} - (y_i)^{a-1} < 0$.

Remarques

A différence de revenus $y_j - y_i$ constante, $|dA(a)|$ est d'autant plus élevée que les revenus sont faibles car $y \to -y^{a-1}$ est une fonction concave : A(a) accorde un poids plus important à l'inégalité parmi les pauvres.

- On peut retrouver les résultats précédents en notant que :

  * si a>0, A(a) varie comme $-[1-A(a)]^a$, qui est de la forme $-\sum_{k=1}^{n} u(y_k)$ *avec* $u(y) = \frac{1}{n} y^a$, et donc $u''(y) < 0$;

  * si a<0, A(a) varie comme $[1-A(a)]^a$, qui est de la forme $-\sum_{k=1}^{n} u(y_k)$ *avec* $u(y) = -\frac{1}{n} y^a$, et donc $u''(y) < 0$.

IV.4 Indice de Theil

IV.4.1 Entropie d'une distribution de revenus

De façon générale, l'**entropie d'une distribution de probabilités** $(p_1 \dots p_k \dots p_n)$ sur un ensemble fini à n éléments (donc vérifiant $p_k \geq 0$, $\sum_{k=1}^{n} p_k = 1$) est définie par :

$$H = \sum_{k=1}^{n} p_k \, Log\left(\frac{1}{p_k}\right)$$

avec la convention 0 log $\left(\frac{1}{0}\right) = 0$

Cette notion d'entropie est dérivée de la théorie de l'information.

L'entropie est maximale, et vaut Log n, si toutes les probabilités sont égales (voir démonstration en annexe).

L'entropie est minimale, et vaut 0, si toutes les probabilités sont nulles sauf une : $\exists i \;\; p_i = 1$

(en effet, dès que $\exists$ k tel que $0 < p_k < 1$, *alors* $H > 0$).

Par analogie, l'**entropie d'une distribution de revenus** $(x_1 \ldots x_k \ldots x_n)$ est définie par :

$$H = \sum_{k=1}^{n} \frac{x_k}{n\bar{x}} Log\left(\frac{n\bar{x}}{x_k}\right)$$

IV.4.2 Définition de l'indice de Theil

L'indice de Theil est égal à la variation d'entropie entre la situation parfaitement égalitaire (revenus tous égaux) et la situation réelle.

$$T = H_{max} - H = Log\, n - \sum_{k=1}^{n} \frac{x_k}{n\bar{x}} Log\left(\frac{n\bar{x}}{x_k}\right)$$

soit :

$$\boxed{T = \sum_{k=1}^{n} \frac{x_k}{n\bar{x}} Log\left(\frac{x_k}{\bar{x}}\right) = \frac{1}{n} \sum_{k=1}^{n} y_k \; Log\; y_k}$$

(avec la convention $y_k \; Log \; y_k = 0 \;\; si \;\; y_k = 0$).

Remarque

$$T = \lim_{a \to 1} \frac{1 - [1 - A(a)]^a}{1 - a}$$

où A(a) est l'indicateur d'Atkinson (§ IV.3) (voir démonstration en annexe).

Valeurs minimale et maximale

D'après le paragraphe précédent :

- $T_{min} = 0$: situation égalitaire
- $T_{max} = Log\ n$: tous les revenus sont nuls sauf un.

IV.4.3 Conséquence d'un transfert

L'indice de Theil est de la forme $-\sum_k u(y_k)$, avec $u(y) = -\frac{1}{n}\, y\, Log\, y$

$u''(y) = -\frac{1}{ny} < 0$: la condition de Pigou-Dalton-Lorenz est vérifiée.

La variation de T consécutive à un transfert infinitésimal h d'un riche j vers un pauvre i vaut :

$$dT = \frac{h}{n}(Log\ y_i - Log\ y_j)$$

Remarque : $u''(y)$ est croissante, donc T accorde plus d'importance à l'inégalité parmi les pauvres qu'à l'inégalité parmi les riches.

IV.5 Coefficient de variation

IV.5.1 Définition

Le coefficient de variation est le rapport de l'écart-type de la distribution des x_k à sa moyenne $\overline{x}$, ou encore l'écart-type de la distribution des y_k :

$$CV = \frac{1}{\overline{x}}\left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2\right)^{1/2} = \left(\frac{1}{n}\sum_{k=1}^{n}(y_k - 1)^2\right)^{1/2}$$

Autres expressions

$$CV = \left(\frac{1}{n} \sum_k (y_k)^2 - 1 \right)^{1/2} = \left(\frac{1}{2n^2} \sum_k \sum_{k'} \left(y_k - y_{k'} \right)^2 \right)^{1/2}$$

Valeurs minimale et maximale

- $CV_{min} = 0$: tous les revenus sont égaux.

- $CV_{max} = \sqrt{n-1}$: tous les revenus sont nuls sauf un (voir démonstration en annexe).

IV.5.2 Conséquence d'un transfert

$(CV)^2$ est de la forme $-\sum_k u(y_k)$, avec $u(y) = -\frac{1}{n}(y-1)^2$, $u'(y) = \frac{2}{n}(y-1)$

La variation de CV consécutive à un transfert infinitésimal h d'un riche j vers un pauvre i vaut :

$$d(CV) = CV \frac{h}{n} (y_i - y_j)$$

La condition de Pigou-Dalton-Lorenz est vérifiée.

Remarque : à différence de revenu $y_j - y_i$ constante, d(CV) est indépendante du niveau des revenus (u"(y) est constante).

V. Désagrégation des indices

On suppose la population P décomposée en M sous-populations, ou strates, $P_1 \dots P_m \dots P_M$.

On note, pour chaque strate P_m :

- n_m l'effectif

- x_{mj} le revenu du j-ième individu de la strate $(j = 1 \dots n_m)$

- $\bar{x}_m = \frac{1}{n_m} \sum_{j=1}^{n_m} x_{mj}$ le revenu moyen.

V.1 *Indice de Theil*

On peut décomposer l'indice de Theil de la façon suivante :

$$T = \frac{1}{n} \sum_{m=1}^{M} \sum_{j=1}^{n_m} \frac{x_{mj}}{\bar{x}} Log\left(\frac{x_{mj}}{\bar{x}}\right) = \frac{1}{n} \sum_{m=1}^{M} \frac{\bar{x}_m}{\bar{x}} \left[\sum_{j=1}^{n_m} \frac{x_{mj}}{\bar{x}_m} Log\left(\frac{x_{mj}}{\bar{x}_m}\right) + \left(\sum_{j=1}^{n_m} \frac{x_{mj}}{\bar{x}_m} \right) Log\left(\frac{\bar{x}_m}{\bar{x}}\right) \right]$$

soit :

$$\boxed{T = \sum_{m=1}^{M} \frac{n_m \bar{x}_m}{n \bar{x}} T_m + \sum_{m=1}^{M} \frac{n_m \bar{x}_m}{n \bar{x}} Log\left(\frac{\bar{x}_m}{\bar{x}}\right)}$$

avec $T_m = \frac{1}{n_m} \sum_{j=1}^{n_m} \frac{x_{mj}}{\bar{x}_m} Log\left(\frac{x_{mj}}{\bar{x}_m}\right)$ = indice de Theil dans la strate P_m .

$\frac{n_m \bar{x}_m}{n \bar{x}}$ représente la part des revenus des individus de la strate P_m .

Le premier terme est la moyenne des indices de Theil calculés à l'intérieur de chaque strate, pondérés par les poids (en termes de revenus) des strates : on l'appelle **indice de Theil intra-strates**.

Le deuxième terme s'interprète comme l'indice de Theil de la distribution des revenus moyens des strates $\bar{x}_1 \ldots \bar{x}_m \ldots \bar{x}_M$, ces revenus étant affectés de poids n_m / n proportionnels aux effectifs des strates ($\bar{x}$ est la moyenne de cette distribution) : on l'appelle **indice de Theil inter-strates**.

On peut donc écrire :

$$\boxed{T = T_{\text{intra}} + T_{\text{inter}}}$$

Remarque : $T_{inter} = 0 \Leftrightarrow \frac{\overline{x}_m}{\overline{x}} =$ constante $\Leftrightarrow$ les revenus moyens par strate sont égaux.

V.2 Ecart moyen de logarithmes

On peut décomposer l'écart moyen des logarithmes (défini au § IV.3.1) de la façon suivante :

$$I = -\frac{1}{n}\sum_{m=1}^{M}\sum_{j=1}^{n_m} Log\left(\frac{x_{mj}}{\overline{x}}\right) = -\frac{1}{n}\sum_{m=1}^{M}\left[\sum_{j=1}^{n_m} Log\left(\frac{x_{mj}}{\overline{x}_m}\right) + n_m\, Log\left(\frac{x_m}{\overline{x}}\right)\right]$$

soit :

$$\boxed{I = \sum_{m=1}^{M}\frac{n_m}{n} I_m + \sum_{m=1}^{M} -\frac{n_m}{n} Log\left(\frac{x_m}{\overline{x}}\right)}$$

avec $I_m = -\frac{1}{n_m}\sum_{m=1}^{n_m} Log\left(\frac{x_{mj}}{\overline{x}_m}\right)$ = écart moyen des logarithmes dans la strate P_m.

Le premier terme est la moyenne des écarts moyens des logarithmes calculés à l'intérieur de chaque strate, pondérés par les poids (en termes d'effectifs) des strates : on l'appelle **écart moyen des logarithmes intra-strates**.

Le deuxième terme s'interprète comme l'écart moyen des logarithmes de la distribution des revenus moyens des strates $\overline{x}_1 \ldots \overline{x}_m \ldots \overline{x}_M$, ces revenus étant affectés de poids n_m / n proportionnels aux effectifs des strates ($\overline{x}$ est la moyenne de cette distribution) : on l'appelle **écart moyen des logarithmes inter-strates**.

On peut donc écrire :

$$\boxed{I = I_{intra} + I_{inter}}$$

Démonstration IV.1.2 (b)

Population des riches $P_r = \{k,\ x_k \geq \bar{x}\} = \{k,\ y_k \geq 1\}$

Population des pauvres $P_p = \{k,\ x_k < \bar{x}\} = \{k,\ y_k < 1\}$

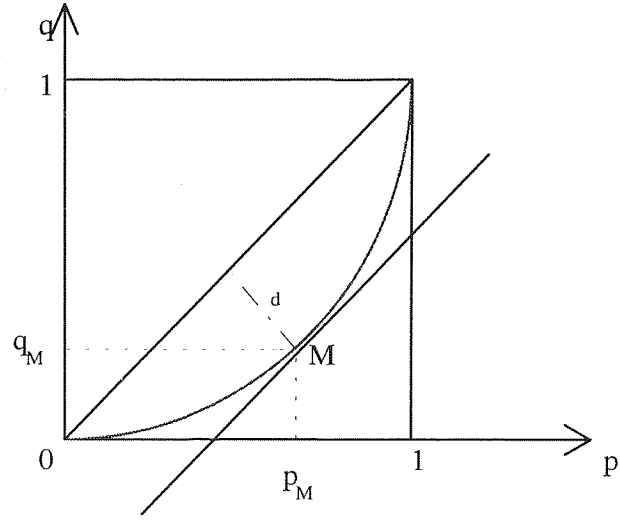
On note n_r *et* n_p les effectifs de P_r *et* P_p, $\bar{x}_r$ *et* $\bar{x}_p$ les revenus moyens, $\bar{y}_r$ *et* $\bar{y}_p$ les revenus relatifs moyens.

On a : $n = n_r + n_p \qquad \bar{x} = n_r\,\bar{x}_r + n_p\,\bar{x}_p \qquad n = n_r\,\bar{y}_r + n_p\,\bar{y}_p$

$$D = \frac{1}{n}\left[\sum_{k\in P_r}(y_k - 1) + \sum_{k\in P_p}(1 - y_k)\right] = \frac{1}{n}\left[\left(n_r\,\bar{y}_r - n_p\,\bar{y}_p\right) + n_p - n_r\right] = 2\left(\frac{n_r\,\bar{y}_r}{n} - \frac{n_r}{n}\right)$$

Démonstration IV.1.2 (c)

La démonstration est simple dans le cas d'une variable continue.



Soit M le point de la courbe de Lorenz où la tangente est parallèle à la bissectrice : c'est le point où la distance à cette bissectrice est maximale (en raison de la convexité de la courbe).

L'abscisse p_M de ce point est telle que $\frac{1}{m} F^{-1}(p_M) = 1$, i.e. $p_M = F(m) = \int_0^m f(u)\,du$

p_M est donc la proportion d'individus "pauvres" (de revenu inférieur à la moyenne).

L'ordonnée q_M de ce point est :

$$q_M = \frac{1}{m} \int_0^{p_M} F^{-1}(v)\,dv = \frac{1}{m} \int_0^m u\, f(u)\,du$$

q_M est la proportion de la masse totale des revenus détenue par les individus "pauvres".

On a alors :

$$\sqrt{2}\,d = p_M - q_M = \frac{1}{m} \int_0^m (m-u)\, f(u)\,du$$

$$et\ D = \frac{1}{m} \int_0^m (m-u)\, f(u)\,du + \frac{1}{m} \int_m^\infty (u-m)\, f(u)\,du = 2 \times \frac{1}{m} \int_0^m (m-u)\, f(u)\,du = 2\sqrt{2}\,d$$

Démonstration IV.2.2

La variance des logarithmes est de la forme $-\sum_k u(y_k)$, avec $u(y) = -\frac{1}{n}\left[Log\left(\frac{y}{g}\right)\right]^2$

$u'(y) \approx -\frac{2}{ny} Log\left(\frac{y}{g}\right)$, $u''(y) \approx -\frac{2}{ny^2}\left[1 - Log\left(\frac{y}{g}\right)\right]$ (en négligeant la variation de g, qui dépend également de y).

On a donc :

$u''(y) < 0$ *si* $y/g < e$

$u''(y) > 0$ *si* $y/g > e$

Démonstration IV.3.1

On a : $\frac{y_k}{n} \le 1 \quad \forall k$, donc : $\sum_{k=1}^{n}\left(\frac{y_k}{n}\right)^a \ge \sum_{k=1}^{n}\frac{y_k}{n} = 1$, avec égalité $\Leftrightarrow$ $y_k = 0 \text{ ou } 1 \ \forall k$

$Donc : \frac{1}{n}\sum_{k=1}^{n}\left(\frac{y_k}{n}\right)^a \ge n^{a-1}$, $et \quad A(a) \le 1 - n^{\frac{a-1}{a}}$, avec égalité $\Leftrightarrow y_n = 1 \ et \ y_k = 0 \ si \ k \ne n$

Démonstration IV.3.2

$Si\ a \ne 0,\ on\ a : (1-A)^a = \frac{1}{n}\sum_{k=1}^{n}(y_k)^a,\ d'où : \ -a(1-A)^{a-1}\,dA = \frac{1}{n}\left[-a(y_j)^{a-1}\,h + a(y_i)^{a-1}\,h\right]$

$soit\ dA(a) = \frac{(1-A)^{1-a}h}{n}\left((y_j)^{a-1} - (y_i)^{a-1}\right)$

$Si\ a = 0,\ on\ a : \ -Log\,[1-A(0)] = -\frac{1}{n}\sum_{k=1}^{n} Log\ y_k,\ d'où : [1-A(0)]^{-1}\,dA(0) = -\frac{1}{n}\left(-\frac{h}{y_j} + \frac{h}{y_i}\right)$

$soit\ dA(0) = \frac{[1-A(0)]h}{n}\left(\frac{1}{y_j} - \frac{1}{y_i}\right)$

Démonstration IV.4.1

Il faut minimiser H sous la contrainte $\sum_k p_k = 1$

Lagrangien L = $-\sum_k p_k\ Log\ p_k - \lambda\left(\sum_k p_k - 1\right)$

$\frac{\partial L}{\partial p_k} = -Log\ p_k - 1 - \lambda = 0$

$\Rightarrow Log\ p_k = \text{constante} \Rightarrow p_k = \frac{1}{n}\forall k$

Démonstration IV.4.2

$Si\ a \approx 1 : (y_k)^a \approx y_k[1 + (a-1)Log\ y_k]$

$$\frac{1}{n}\sum_k (y_k)^a \approx \frac{1}{n}\sum_k y_k + \frac{a-1}{n}\sum_k y_k \, Log\, y_k = 1 + (a-1)T$$

$$\Rightarrow (1-a)T = 1 - \frac{1}{n}\sum_k (y_k)^a = 1 - [1 - A(a)]^a$$

Démonstration IV.5.1

$$(CV)^2 + 1 = \frac{1}{n}\sum_k (y_k)^2 = n\sum_k \left(\frac{y_k}{n}\right)^2 \leq n\sum_k \frac{y_k}{n} = n \text{ , avec égalité } \Leftrightarrow y_k = 0 \text{ ou } 1 \; \forall k$$

D'où : $CV \leq \sqrt{n-1}$, avec égalité $\Leftrightarrow y_n = 1$ *et* $y_k = 0$ *si* $k < n$

ESTIMATION DE LA VARIANCE DU COEFFICIENT DE GINI MESURÉ PAR SONDAGE

Jean-Claude Deville

Le coefficient de GINI mesure la dispersion d'une variable numérique positive dans une population. Il est fréquemment utilisé dans les Instituts de Statistique notamment pour évaluer les disparités et les inégalités de revenus. C'est, par ailleurs, un exemple caractéristique de fonctionnelle "fortement" non linéaire, qu'on ne peut en aucun cas assimiler à une fonction de totaux. Il paraît donc a priori difficile d'évaluer la précision d'un tel indice quand il est calculé à partir des résultats d'une enquête par sondage car la théorie habituelle (WOODRUFF[10]) ne s'applique pas. Cette note s'attaque à un problème qui présente un réel intérêt tant pratique que théorique. On y propose une solution qui résout le problème par des techniques appropriées de linéarisation et qui le ramène à un calcul classique d'estimation de variance d'un total pour l'estimateur de HORVITZ-THOMPSON. La variance de l'indice de GINI peut donc être évaluée à l'aide d'un logiciel de calcul de précision de type standard comme celui auquel travaille l'équipe de l'Unité Méthodes Statistiques.La solution qui est proposée est assez originale bien qu'on puisse en trouver la source dans [1] : on y montre que l'indice de GINI peut-être vu comme un ratio d'estimateurs et qu'une méthode d'estimation de variance peut-être dérivée de cette idée. Toutefois, dans [2], on constate que la technique du ratio donne dans des simulations des résultats extraordinairement mauvais, la variance étant surestimée de 30 à 300 fois. On montre ici qu'il y avait un étrange artefact dans la façon dont on attaquait le problème dans ces articles et qu'une analyse plus fine de la notion de linéarisation permet d'arriver à un résultat correct.On commencera, en prime, par réviser les formulations possibles du coefficient de GINI. On passera ensuite à la question de son estimation dans les sondages complexes, puis à celui de l'estimation de la variance. Une simulation montrera ensuite que l'expérience est en accord très honorable avec les résultats théoriques.

1 - Coefficient de GINI

Comme beaucoup, sans doute, une des seules choses que je savais du coefficient de GINI est ce qu'on trouve dans le livre de G. CALOT [3]. On définit, pour une variable réelle positive Z les objets suivants :

- Sa fonction de répartition F(t) = Fréquence de $Z \leq t$

- Sa moyenne $M = \int t\, dF(t)$

- la fonction de répartition

$$G(t) = \frac{1}{M}\int^{t} z dF(z)$$

- La courbe de concentration de LORENZ définie comme l'ensemble des points :

$$\begin{cases} x(t) = G(t) \\ y(t) = F(t) \end{cases}$$

S'il y a des discontinuités, la courbe de LORENZ est l'enveloppe convexe de ces points auxquels on ajoute (0,0). Cette courbe est convexe, contenue dans le carré $[0,1]\times[0,1]$ passe par $[0,0]$ *et* $[1,1]$. Par suite elle se situe au dessus de la diagonale du carré.

L'indice de GINI est alors le double de la surface comprise entre la courbe et cette diagonale. Il est proche de 0 si Z est presque constante. Il est proche de 1 si F(0) tend vers 1. Cet indice est souvent utilisé pour décrire les inégalités de salaires. Une valeur proche de 1 décrit donc une situation très inégalitaire où une petite fraction de la population dispose d'une grosse partie de l'ensemble des revenus.

La littérature nous fournit d'autres expressions plus commodes sur le plan analytique et permettant de calculer l'indice autrement que graphiquement en comptant le nombre de carreaux sous la courbe. Par ailleurs, le pont entre les diverses expressions de l'indice est rarement indiqué dans les ouvrages (en particulier aucune des références citées ici ne le fait !) bien que (ou parce que ?) ce soit assez peu évident.

Commençons par cela .

La formule d'intégration par partie pour les intégrales de Stieltjes nous sera utile. Rappelons celle-ci (qui est immédiate à établir). Soient F et G sont deux fonctions croissantes définies sur $\mathbb{R}^+$ (fonctions de répartition de mesures positives). On les supposera continues à droite. Elles admettent en tout point une limite à gauche qu'on notera F(*t-0*). La quantité $\Delta F(t) = F(t) - F(t-0)$ s'appelle le saut de F en *t*. On posera F(0-0) = 0. On a alors la formule d'intégration par parties (pour a $\leq$ b)

$$F(b)\,G(b) - F(a-0)\,G(a-0) = \int_a^b F(t)\,dG(t) + \int_a^b G(t)\,dF(t) - \sum_{t\in[a,b]} \Delta F(t)\,\Delta G(t).$$

Le troisième terme concerne donc les points où F et G ont des sauts communs et indique une famille sommable [9] de termes positifs.

La définition graphique de l'indice de GINI nous conduit à l'expression analytique :

$$GINI = 2 \int F(t)dG(t) - \sum \Delta F(t)\,\Delta G(t) - 1$$

$$= \frac{2}{M} \int t\ F(t)dF(t) - \sum_t t\ \Delta F(t)^2 - 1 \qquad (2)$$

Mais l'intégration par partie nous donne :

$$\int FdG = \left[FG\right]_{0-0}^{+\infty} - \int GdF + \sum_t \Delta G(t)\ \Delta\ F(t)$$

d'où :

$$GINI = \int\ F(t)dG(t) - \int\ G(t)dF(t)$$

$$= \frac{1}{M}\left[\int\ tF(t)dF(t) - \int\ dF(t)\int^t udF(u)\right]$$

$$= \frac{1}{M}\left[\int\ dF(t)\int^t\ (t-u)\,dF(u)\right]$$

et donc, par symétrie :

$$GINI = \frac{1}{2M} \iint\ |t-u|\,dF(t)\,dF(u) \qquad (3)$$

Remarque *: La formule (2) fait apparaître, dans le cas continu en tous cas, la moyenne du supremum de deux variables indépendants suivant la loi F. J'ignore ce que cette remarque apporte de plus qu'un élément de cocasserie.*

2 - Cas d'une population finie

Traduisons les résultats précédents dans le cas d'une population finie d'effectif N. La variable d'intérêt z prend des valeurs z_k *(pour* $k = 1$ *à* N*)* et on supposera que $z_k \leq z_{k+1}$.

La traduction de la formule (3) est alors :

$$GINI = \frac{1}{2} \frac{\sum_{k=1}^{N} \sum_{\ell=1}^{N} |Z_k - Z_\ell|}{N \sum_{i=1}^{N} Z_i}$$

On constate que GINI = 0 quant tous les z_k sont égaux. GINI est maximum quand $z_k=0$ pour k = 1 à N-1 et $z_N>0$. On trouve alors la valeur de $1-\frac{1}{N}$.

Passons maintenant à la formule (2).
Si nous supposons $z_k < z_{k+1}$ pour tout k on aura alors F(t) = k/N pour t dans l'intervalle $\left[z_k, z_{k+1}\right[$ (avec la convention $z_0 = 0$ et $z_{N+1} = +\infty$). La formule de l'indice de GINI devient alors :

$$GINI = \frac{2 \sum_{k=1}^{N} k z_k}{N \sum_{k=1}^{N} z_k} - \frac{\sum_{k=1}^{N} z_k}{N^2 M} - 1 = \frac{2 \sum_{k=1}^{N} k z_k}{N \sum_{k=1}^{N} z_k} - \frac{1}{N} - 1$$

$$= \frac{\sum_{k=1}^{N} (2k-1) z_k}{N \sum_{k=1}^{N} z_k} - 1 \qquad (2')$$

Cette forme est particulièrement utile car elle fait apparaître l'indice comme le ratio de deux statistiques assez simple de la population finie : le total au dénominateur et le total de $y_k = \left(\frac{2k-1}{N} - 1\right) z_k$ au numérateur. Cette dernière pose toutefois un problème, car, généralement, le rang k des unités observées est inconnu.

On remarquera, par ailleurs, que la forme (2') reste vraie si plusieurs unités k prennent la même valeur y_k (le classement des ex-aequo étant alors arbitraire). On s'en convaincra soit en remarquant la stabilité par passage à la limite de (2'), soit par un raisonnement direct. Au cas où la valeur z comporte q ex-aequo, l'élément de la formule (2) correspondant à cette valeur s'écrit (à un facteur près) :

$$2z \frac{i}{N} \cdot \frac{q}{N} - z\left(\frac{q}{N}\right)^2 = \frac{z}{N^2}\left(2iq - q^2\right)$$

Or :

$$2iq - q^2 = i^2 - (i-q)^2 = \sum_{r=0}^{q-1} (i-r)^2 - (i-r-1)^2$$

$$= \sum_{r=o}^{q-1} (2(i-r)-1)$$

Ce qui permet de terminer la preuve.

La forme (1) - ou graphique - n'apporte pas véritablement d'information supplémentaire. Le cas fini permet un calcul exact de l'indice défini graphiquement. On doit évaluer l'aire d'une réunion de N trapèzes. Le k^{ème} d'entre eux a pour base z_k/NM et pour demi-hauteur (k-1/2)/N pour k = 1 à N. La surface sous la courbe de LORENZ vaut donc :

$$\sum_{k=1}^{N} \frac{k-\frac{1}{2}}{N} \cdot \frac{z_k}{NM}$$

D'où la formule (2'). Si plusieurs unités k prennent la même valeur z_k il est facile de voir que le calcul précédent reste valide.

3 - Estimation à partir d'une données d'enquêtes par sondage dans le cas où le rang est observable

Commençons par le cas où le rang de classement i est observable. Bien que cela ne soit pas chose courante, voici un exemple qu'on peut imaginer. On veut calculer la dispersion des temps mis par les coureurs du Tour de France à effectuer les étapes successives (vite dit ou veut connaître "scientifiquement" les plus sélectives !). Pour cela on tire avant le départ un échantillon s de coureurs k dont on relèvera, à la fin de chaque étape, le classement i_k et le temps réalisé z_k (pour plus de réalisme ce temps pourrait être l'écart entre le temps d'arrivée du premier et le délai au delà duquel les coureurs sont éliminés).

En notant w_k les poids attribué dans l'enquête, l'indice de GINI sera estimé par un ratio :

$$\hat{GINI} = \frac{\sum_s (2i_k - 1) z_k w_k}{N \sum_s z_k w_k} - 1 \qquad (3-1)$$

Éventuellement, si N n'est pas connu ou si la somme des poids ne vaut pas N on remplace N par $\hat{N} = \sum_s w_k$. La variance de GÎNI s'estime alors comme celle d'un ratio - si N est connu - ou par une formule standard de linéarisation si N est remplacé par son estimateur.

Autrement dit, si on sait calculer (par exemple par un programme standardisé) la variance estimée d'un total de la forme $\hat{Y} = \sum_k w_k y_k$, on remplace y_k dans le calcul par la variable artificielle:

$$U_k = \frac{1}{\hat{Z}}\left(\frac{2i_k - 1}{N} - \left(G\hat{I}NI + 1\right)\right) z_k$$

dans le cas ou N est connu. Dans le cas ou N est estimé on prendra :

$$u_k' = \frac{1}{\hat{N}\,\hat{Z}}\left[\left(2i_k - 1 - \left(G\hat{I}NI + 1\right)\hat{N}\right) z_k - A\right] \quad ,$$

$$avec \qquad A = \sum_s \left(2i_k - 1\right) z_k w_k \Big/ \sum_s w_k$$

Le calcul se ramène donc à une recodification élémentaire.

4 - Estimation dans le cas où le rang n'est pas observable

Malheureusement on ne dispose pas, en général du rang i_k de l'unité k. Celui-ci doit être estimé à partir des données, en se basant sur une estimation $\hat{F}(z)$ de la fonction de répartition de Z.

Cette estimation de $\hat{F}$ peut se faire de diverses façons. Le plus simple et le plus naturel consiste à utiliser les estimateurs de HORVITZ-THOMPSON des proportions de z_k inférieurs ou égaux à z et cela pour tout z. On prend classiquement donc :

$$\hat{F}(z) = \sum_{k:\, z_k \le z} 1/\pi_k \Big/ \sum_{k \in s} 1/\pi_k$$

Ceci posé, les poids de H.T peuvent être remplacés par n'importe quel autre système de poids associé à des performances connues des estimateurs linéaires qu'ils définissent. La fonction $\hat{F}(z)$ est constante par intervalles, continue à droite et à des sauts pour toutes les valeurs z_k prises par l'échantillon. Pour d'autres principes d'estimation (voir par exemple CHAMBERS et DUNSTAN [5]) cette particularité se maintient. En fait, c'est le rang de z_k que nous voulons estimer, ce qui est un peu malheureux car nous avons justement un saut de $\hat{F}$ en z_k et nous pouvons hésiter entre $\hat{F}(z_k)$ *et* $\hat{F}(z_k - 0)$ comme estimateur de i_k/N. Il est raisonnable d'utiliser la moyenne de ces deux valeurs :

$$\tilde{F}(z_k) = \frac{1}{2}\left(\hat{F}(z_k) + \hat{F}(z_k - 0)\right) = \hat{F}(z_k) - \frac{1}{2}\frac{w_k}{\hat{N}}$$

Ce choix correspond à une fonction de répartition $\tilde{F}$ obtenue par un léger lissage. Graphiquement il correspond à utiliser l'estimateur $\tilde{F}$ dont la courbe représentative est obtenue en joignant les milieux les "contremarches" adjacentes de la courbe représentative de $\hat{F}$. L'inconvénient de cet estimateur est de ne pas être défini (sauf convention particulière) à l'extérieur de l'intervalle $\left[min\,(z_k),\ max\,(z_k)\right]$

La quantité $2i_k - 1$ de (3-1) se trouve donc estimée naturellement par $2\,\hat{N}\,\tilde{F}_k - 1$ ce qui conduit à la forme suivante de l'estimateur du GINI:

$$\hat{GINI} = \frac{\sum_s \left(2\hat{N}\,\tilde{F}_k - 1\right) z_k\, w_k}{\hat{N} \sum_s w_k\, z_k}$$

$$= \frac{\sum_s 2\hat{N}\,\tilde{F}_k\, z_k\, w_k}{\hat{N} \sum_s w_k\, z_k} - \left(1 + \frac{1}{\hat{N}}\right) \qquad (4-1)$$

Dans le cas d'un sondage aléatoire simple (ou plus généralement, d'un sondage de taille fixe à probabilité égales - où N est connu), cette expression prend la forme :

$$\hat{GINI} = \frac{\sum_s (2j_k - 1)\, z_k}{n \sum_s z_k} - \left(1 + \frac{1}{N}\right) \qquad (4-2)$$

où j_k est le rang (de 1 à n) de l'unité k dans s.

5 - Considérations globales de biais et de variance

Les estimateurs (4-1) ou même (4-2) ont une forme de ratio et n'ont aucune raison d'être sans biais. Dans les applications concrètes, par ailleurs, il est à peu près évident que la variabilité du terme en $\frac{1}{\hat{N}}$ de (4-1) peut-être négligée (le terme lui-même peut l'être !). Le dénominateur du ratio est une quantité simple. Seul le numérateur demande un examen préliminaire un peu précis. Mettons (4-1) sous la forme :

$$\hat{GINI} = \frac{\sum_s 2\,\tilde{F}_k z_k\, p_k}{\sum_s z_k\, p_k} - 1 \qquad où \qquad p_k = w_k / \hat{N}$$

Le remplacement de $\tilde{F}$ *par* $\hat{F}_k$ produit une variation de l'indice égale à :

$$\sum_s z_k\, p_k^2 \Big/ \sum_s z_k\, p_k$$

qui est de l'ordre de *1/n* car p_k est lui même de l'ordre de *1/n*.

Le biais généré par ce remplacement sera donc en général négligeable. Son incidence sur l'erreur quadratique moyenne se traduira par un terme en $1/n^2$ qu'on ne cherchera pas à capturer.

Par ailleurs, le numérateur lui-même n'estime aucune quantité simple sans biais ! Sans reprendre toute l'analyse faite dans [1], examinons une seconde ce qui se passe dans le cas du sondage aléatoire simple où ce numérateur vaut :

$$NUM = \frac{1}{n^2} \sum_s \left(2j_k - 1\right) z_k = \frac{1}{n^2} \sum_U \varepsilon_k z_k \left(2 \sum_{\ell=1}^{k} \varepsilon_\ell - 1 \right)$$

où ε_k est la variable d'appartenance à l'échantillon $\left(\varepsilon_k = 1 \text{ si } k \in s \text{ et } 0 \text{ sinon}\right)$.

On voit alors que :

$$E(NUM) = \frac{1}{n^2} \sum_{k=1}^{N} z_k \left(2k. \frac{n(n-1)}{N(N-1)} - \frac{n}{N} \right)$$

$$= \frac{1}{N-1} \sum_{k=1}^{N} \frac{2k}{N} z_k - \frac{1}{n} \left(\frac{1}{N} \left(\sum_1^N \frac{2k}{N-1} + 1 \right) z_k \right)$$

On découvre donc de nouveau un biais de l'ordre de 1/n. Pour des plans et des estimateurs plus généraux on pourrait montrer qu'on aurait aussi un biais de cet ordre de grandeur. Ce second biais a le même genre d'incidence sur l'écart quadratique moyen que celui qui a déjà été évoqué.

Cette analyse peut-être résumée par la proposition suivante :

Résultat : La quantité

$$\hat{GINI}^* = \frac{\sum_s 2\, z_k\, \hat{F}_k\, p_k}{\sum_s z_k\, p_k} - 1 \qquad (5-3)$$

est un estimateur de l'indice de GINI qui a, à des termes en $\frac{1}{n}$ près, le même biais que l'estimateur (4-1). De plus, il a le même écart quadratique moyen à des termes en $1/n^2$ près.

C'est, donc sur cette quantité que nous allons travailler maintenant.

6 - Ébauche de linéarisation pour GINI*

Posons $\hat{R} = 2\left(1 + \hat{GINI}^*\right) = NUM / \hat{M}$.

En vertu des règles générales de linéarisation et de la confusion habituelle (et justifiée par l'analyse du biais) entre variance et écart quadratique moyen nous pouvons écrire :

$$VAR\left(\hat{R}\right) \simeq \frac{1}{M^2}\left(Var\,(NUM) - 2R\,Cov(NUM, \hat{M}) + R^2\, Var\left(\hat{M}\right)\right)$$

Le traitement de $\hat{M} = \frac{1}{\hat{N}} \sum_s z_k\, w_k$ ne pose pas de problème particulier dès que l'on sait estimer la variance de l'estimateur du total de Z, $\sum_s z_k\, w_k$. L'estimateur de variance utilise la variable linéarisée $\left(z_k - \hat{M}\right)/\hat{N} = z_k'$ comme le montre un calcul élémentaire.

En revanche, comme nous allons le voir, le traitement de NUM demande plus de finesse. En effet, on a :

$$NUM = \sum_s z_k \hat{F}_k p_k$$

On pourrait donc raisonner sommairement en disant que $z_k \hat{F}_k$ estime $z_k F_k$ et qu'on peut traiter le problème comme au paragraphe 3. C'est ce qui est fait dans [2], et conduit à la conclusion étonnée que la variance estimée est de 10 à 300 fois trop forte.

En fait, c'est négliger la covariance entre $\hat{F}_k$ et le poids p_k (aléatoire !!) attribué à l'unité k. Comme, de plus, ces covariances sont sommées sur toute la population on obtient au final un biais d'ordre fini (même pour de très gros échantillons) c'est-à-dire, pour parler clair, un calcul parfaitement faux.

Nous allons, dans la suite, montrer que la variance de NUM est, à des termes négligeables près), celle du total d'une variable artificielle x_k :

$$VAR\,(NUM) \simeq Var\left(\sum_s x_k\, w_k\right).$$

Par suite on aura :

$$Cov\left(NUM, \hat{M}\right) \simeq Cov\left(\sum_s x_k w_k, \sum_s z_k' w_k\right)$$

Il en résultera, d'après (6-1), que :

$$Var\ \hat{R} \simeq \frac{1}{M^2} Var\left(\sum_s \left(x_k - Rz_k'\right) w_k\right)$$

$$= Var\left(\sum_s r_k w_k\right) avec\ r_k = \frac{1}{M}\left(x_k - Rz_k'\right)$$

Ceci marquera la fin de (presque) tous nos problèmes avec la variance de l'indice de GINI.

7 - Linéarisation de la fonctionnelle NUM

Notons qu'on peut écrire :

$$NUM = \sum_s z_k \hat{F}_k p_k = \int_{IR} z\ \hat{F}(z)\ d\hat{F}(z) \qquad (7)$$

avec $\hat{F}$ fonction de répartition estimée de la variable Z. Cette quantité est prise pour estimateur de $\int_{IR} z\ F(z)\ dF(z)$, de sorte qu'on est amené à examiner la quantité :

$$D = \int z\ \hat{F}(z)\ d\hat{F}(z) - \int z\ F(z)\ dF(z)$$

$$= \int z\ F(z)\ d\left[\hat{F} - F\right](z) + \int z\left(\hat{F}(z) - F(z)\right) dF(z) + \int z\left(\hat{F}(z) - F(z)\right) d\left[\hat{F} - F\right]$$

Traitons le dernier terme ; $Max_z\left(\hat{F}(z) - F(z)\right)$ est une variable aléatoire qui est en probabilité de l'ordre de $1/\sqrt{n}$. Pour de gros échantillons le dernier terme sera donc négligeable vis-à-vis du premier.

Examinons maintenant le second terme qui peut s'écrire.

$$\int\left(\hat{F}(z) - F(z)\right) dG(z) \quad avec \quad G(z) = \int^z u\, dF(u).$$

On peut maintenant utiliser l'intégration par partie :

$$\int\left(\hat{F} - F\right) dG = \left[G\left(\hat{F} - F\right)\right]_o^\infty - \int G\, d\left[\hat{F} - F\right] + \sum \Delta G.\Delta\left(\hat{F} - F\right)$$

Regardons le troisième terme :

$$\sum \Delta G\ \Delta\hat{F} = \sum_s \frac{z_k}{N}.\, p_k = \frac{1}{N}\hat{M}$$

$$\sum\ \Delta G\ \Delta F = \sum_U \frac{z_k}{N}.\frac{1}{N} = \frac{1}{N}M$$

Le troisième terme vaut donc $\frac{1}{N}\left(\hat{M}-M\right)$ et peut donc être considéré comme négligeable.

Le premier est nul car $\hat{F}(o-o)=F(o-o)=0$ *et* $\hat{F}(\infty)=F(\infty)=1$.

On obtient donc que :

$$D \simeq \int \left(z\ F(z)-G(z)\right)\ d\left(\hat{F}(z)-F(z)\right)$$

à des quantités négligeables près.

Il en résulte que la variance de D (ainsi que les covariances de D avec d'autres variables) est donnée au premier ordre, par la variance de la variable :

$$x_k = z_k\ F\left(z_k\right)-G\left(z_k\right).$$

On remarquera qu'on a :

$$G\left(z_k\right)=F\left(z_k\right) z_k^*$$

où z_k^* est la moyenne de la variable Z pour la population des individus ℓ vérifiant $z_\ell \leq z_k$. On écrira donc :

$$x_k = F_k\left(z_k - z_k^*\right),$$

et nous sommes arrivés au but que nous nous étions fixé.

8 - Estimation de la variance de GINI

Synthétisons les résultats :

- *NUM* se linéarise en $\frac{1}{\mathrm{N}}\left(x_k-\bar{x}\right)$
- $\hat{M}$ se linéarise en $\frac{1}{\mathrm{N}}\left(z_k-\bar{z}\right)$
- $\hat{R}$ se linéarise $\frac{1}{\mathrm{M}}.\frac{1}{\mathrm{N}}\left(F_k\left(z_k-z_k^*\right)-Rz_k-\bar{u}\right)$

avec $\bar{u}$ moyenne de $u_k = F_k\left(z_k-z_k^*\right)-Rz_k$

On a alors :

$$Var\left(\hat{GINI}\right) \simeq Var\left[\left(\sum_s \left(u_k - \overline{u}\right) w_k\right) / 2MN\right]$$

On est donc ramené à l'estimation d'un total pour une variable artificielle bien précise. Malheureusement, celle-ci n'est pas connue sur l'échantillon mais peut être, dans le calcul, remplacée par des estimations sans que cela fasse de dégâts. En effet, l'estimateur de variance calcule une forme quadratique finie :

$$\sum\sum_s A_{k\ell}\, x_k\, x_\ell$$

Si les x_k sont remplacés par des estimations introduisant des erreurs numériques de l'ordre de $1/\sqrt{n}$, l'erreur sur l'estimateur de variance sera également de l'ordre de $1/\sqrt{n}$ au plus, ce qui reste parfaitement honorable.

9 - Procédure de calcul

La chaîne de calcul pour l'estimation de GINI et l'estimation de sa variance procède donc comme suit :

1 - Ordonner les données dans le sens des z_k croissant.

2 - Calculer les cumuls W_k *des* w_k *et* Z_k *des* $z_k\, w_k$.

3 - Calculer les estimateurs ponctuels :

$$* \hat{N} = \sum_s w_k\,, \qquad * \hat{Z} = \sum_s z_k\, w_k$$

$$NUM = \sum_s z_k \frac{W_k}{\hat{N}} w_k$$

$$\hat{R} = NUM / \hat{Z}$$

$$\hat{GINI} = 2\,\hat{R} - 1.$$

4 -Calculer pour tout k la variable "artificielle" :

$$u_k = \frac{1}{\hat{Z}}\left(F_k\left(z_k - \frac{Z_k}{W_k}\right) - \hat{R} z_k\right)$$

5 - Calculer $\overline{u} = \frac{1}{\hat{N}} \sum u_k w_k$.

6 - Calculer la variance de $u_k - \overline{u}$ à l'aide d'un logiciel approprié.

10 - Une petite simulation pour se réconforter

Quand on cherche à établir un résultat de statistique par des voies théoriques, on est souvent conduit à s'interroger sur la validité des approximations qu'on a été amené à faire au cours de l'analyse. Comme en physique, une vérification expérimentale est souvent utile pour se convaincre que tout marche bien comme on le soupçonne (ou comme on le souhaite !). En particulier les résultats de [2] étaient particulièrement intriguants et inquiétants. On a donc réalisé une petite simulation dans l'esprit de cette étude de façon à en avoir le cœur net. Pour des raisons de volume de calcul, et aussi parce que le problème des petits échantillons dans les petites populations est relativement critique pour juger de la validité des approximations, présentons ici les résultats relatifs à une population comptant N = 200 individus dans laquelle on a échantillonné, par sondage aléatoire simple, n = 20 puis 40 individus. On a simulé ainsi 10 000 échantillons indépendants pour établir les résultats.

Le seul point qui pose un réel problème est celui de l'estimation de la fonctionnelle NUM. La simulation qui est présentée dans la suite vérifie que la variance est estimée décemment par linéarisation. Les individus de la population sont numérotés de k=1 à 200, et, à chacun d'eux on associe la variable:

$$z_k = (k/N)^a + 1/a$$

On a donné à a 12 valeurs échelonnées de 0,1 à 12,5 pour simuler des variables de plus en plus concentrées. Vu la définition de ces variables , le rang R_k de l'individu k vaut k . Pour un échantillon s de taille n , on notera $i_{k(}$(=1 à n) le rang de l'individu k dans s . On considérera deux cas: celui où le rang R_k (= k) de l'individu échantillonné est effectivement observé (voir paragraphe 3) et celui où il est estimé à partir de l'échantillon par $\hat{R}_k = \frac{N}{n}(i_k - 0,5)$.

Autrement dit la quantité :

$$NUM = \sum_{k=1}^{N} (R_k - 0,5)\, z_k$$

a été estimée par :

$$\hat{NUM1} = \frac{N}{n} \sum_{k \in s} (R_k - 0{,}5) z_k = \frac{N}{n} \sum_{i=1}^{n} (R_{k_i} - 0{,}5) z_{k_i} \quad (10.1)$$

et :

$$\hat{NUM2} = \frac{N}{n} \sum_{k \in s} \hat{R}_k \; z_k = \frac{N}{n} \sum_{i=1}^{n} \frac{N}{n} (i - 0{,}5) z_{k_i} \quad (10.2)$$

Les résultats sont présentés dans le tableau annexé. En colonne 1 figure la constante a , en colonne 2 la valeur exacte de NUM.

Les colonnes 3 à 5 sont relatives à NÛM1 et présentent respectivement :

(3) : le biais relatif calculé par $\frac{1}{B} \frac{\sum (\hat{NUM1} - NUM)}{NUM} \times 100$

avec B = 10000 simulations et la somme qui porte sur ces 10000 expériences.

(4) :l'écart-type relatif calculé par $\frac{\sqrt{\frac{1}{B} \sum (\hat{NUM1} - NUM)^2}}{NUM} \times 100$.

(5) : l'écart-type relatif estimé. Si $\hat{V}$ est une estimation (parmi les 10000) de la variance de $\hat{NUM1}$, on calcule la quantité $\frac{\sqrt{\frac{1}{B} \sum \hat{V}}}{NUM} \times 100$. Ici, la variance $\hat{V}$ a donc été calculée en utilisant la procédure du paragraphe 3. Comme on pouvait s'y attendre , on ne décèle aucun biais significatif dans l'estimation de NUM; de même , la variance de NÛM est estimée sans biais apparent.

Les colonnes 6, 7, et 8 sont relatives à NÛM2 et donnent aussi le biais relatif, l'écart-type relatif et l'estimation de l'écart-type relatif. L'estimateur de variance utilisé ans cette colonne 8 est celui qui est décrit au paragraphe 9.

On constate un léger biais négatif pouvant atteindre -1,4% pour des variables très concentrées et des échantillons de taille 20 . La variance (ou l'écart-type relatif) est estimée très honorablement ; une légère sous-estimation apparaît lorsque le biais est important ce qui est bien naturel.

La colonne 9 donne "l'estimateur" de variance (ou d'écart-type relatif) basé sur la variable naïve utilisée dans [1] ou [2]. La variance est alors surestimée de 1,5 à 14000 fois quand on passe d'une variable très concentrée (a= 12,5) à une variable très dispersée. L'ordre de grandeur de 300 fois cité dans [2] pour une variable naturelle apparaît comme plausible.

Numériquement cet estimateur est très voisin de celui qui apparaît en colonne 5 , et on peut donc dire qu'en fait il estime la même quantité, c'est à dire la variance de NÛM1 . Ceci a quelque chose de très surprenant : NÛM1 utilise la variable exacte R_k alors que NÛM2 utilise son estimation $\hat{R}_k$; cependant , le deuxième estimateur est bien meilleur !

Quelle est l'explication de ce mystère ?

Je n'ai pas de réponse définitive mais on peut constater qu'on peut écrire :

$$\hat{NUM1} = \int z_k \, F_k \, d\hat{F}_k$$

et $$\hat{NUM2} = \int z_k \, \hat{F}_k \, d\hat{F}_k$$

Si les z_k étaient constants (mais pas les F_k !) , on aurait $\hat{NUM2} = \overline{Z}/2$, avec une variance nulle ,tandis que NÛM1 conserverait une variance positive. D'un autre point de vue , regardons les formules 10-1 et 10-2 valides pour le sondage aléatoire simple .Dans la première les R_{k_i} diffèrent d'un échantillon à l'autre alors que dans la seconde les quantités iN/n ne varient pas ! On doit donc s'attendre à une variance plus grosse dans le premier cas que dans le second.

Il y a la un artefact statistique aussi étonnant que paradoxal , une véritable curiosité !

11 - Conclusion et voie d'avenir

On arrive ainsi à un traitement relativement honorable du problème de la variance du coefficient de GINI estimé sur des données d'enquête . Il n'a recours qu'à des traitements relativement standardisés et est adaptable, en principe à n'importe quelle enquête ayant un plan de sondage complexe. Il évite, en particulier, un traitement sur mesure par rééchantillonnage (Jacknife, Bootstrap, groupes aléatoires, demi-échantillons équilibrés). Les techniques de rééchantillonnage, en effet, ne sont adaptées qu'à certains plans, demandent toujours une programmation relativement spécifique et lourde, et des calculs qui peuvent atteindre un volume prohibitif.

Il se trouve, que, pendant que cette étude était en gestation, j'ai pu avoir accès à un projet de publication de D. BINDER [8]. Celui-ci s'intéresse au calcul de variance de certains indicateurs de dispersion dont l'indice de GINI. En utilisant la technique de linéarisation des équations estimantes, ils parvient, pour l'indice de GINI au même résultat, que celui qui est établi plus haut. Sa technique semble apparemment plus générale que celle que nous avons utilisée. Un examen soigneux des arguments montre que ce n'est qu'une apparence.

En fait, une véritable généralisation de ce qui vient d'être exposé devrait passer par l'utilisation de la notion de fonction d'influence telle qu'on l'utilise en statistique non paramétrique robuste. En modifiant un peu la notion pour tenir compte du contexte des populations finies, on peut arriver à des résultats extrêmement généraux et simples à la fois. Ceux-ci seront (si les petits cochons ne me mangent pas) exposés dans une prochaine étude.

BIBLIOGRAPHIE

[1] "The Estimation of the GINI and the Entropy Inequality Parameters in Finite Populations", Frédérik NYGARD and Arne SANDSTRÖM, *Journal of Official Statisitics*, Vol 1 n° 4, 1985 pp : 399-412

[2] Arne SANDSTRÖM, Jan WRETMAN and Bertil WALDEN, "Variance Estimators Of the GINI COEFFICIENT" In Simple random Sampling, origine douteuse, vraisemblablement GENUS (1987) pp : 41 à 70

[3] Gérard CALOT : *Statistique Descriptive* (DUNOD 1965)

[4] M. FLEURBAEY et S. LOLLIVIER : "Les mesures des inégalités : Abrégé théorique et pratique", Document de Travail du CREST n° 9408 bis (INSEE, 1994)

[5] R. CHAMBERS and R. DUNSTAN : "Estimating Distribution Functions from Survey Data", - *Biometrika* 1986 - vol 73 pp : 597-604

[6] Jean-Michel HOURRIEZ : "La significativité des variations du coefficient de GINI", Note "manuscrite" INSEE (1995)

[7] Ph. TASSI - B. LECOUTRE : *Statistique non-paramétrique et Robustesse,* Economica (1987).

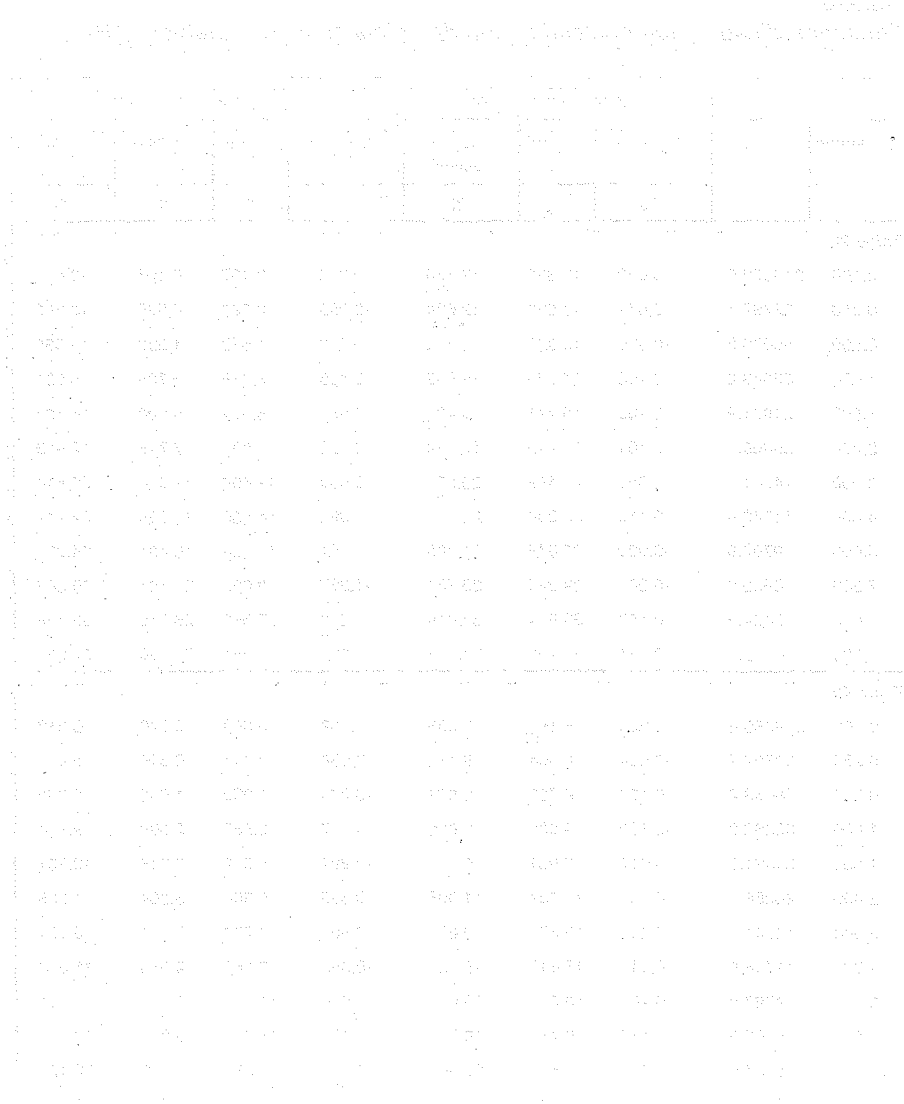
[8] Milorad S. KOVACEVIC and David A. BINDER : "Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach", Document interne de Statistique Canada (1995).

[9] P.R. HALMOS : *An Introduction to Hilbert Spaces and Spectral Multiplicities.*

[10] WOODRUFF, R.S (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate", *Journal of the American Statistical Association* 66, 411-414.

Tableau 1 :
Simulation d'estimation et d'estimation de précision de la variables NUM.

		Le rang est observé			Le rang n'est pas observé			
Constante	Total	biais rel	ect rel	ect rel estimé	biais rel	ect rel	ect rel est	pseudoest
		%	%	%	%	%	%	%
Taille 20								
0.100	219056.6	-0.052	12.325	12.360	-0.018	0.107	0.107	12.617
0.250	97797.7	-0.041	12.599	12.624	-0.083	0.527	0.525	12.886
0.500	56033.2	-0.036	13.308	13.310	-0.218	1.550	1.520	13.538
1.000	33383.2	-0.163	14.736	14.832	-0.495	3.815	3.726	15.031
1.500	24821.9	0.103	16.271	16.408	-0.580	5.974	5.820	16.525
2.000	20066.7	0.101	17.748	17.839	-0.722	7.994	7.718	17.913
3.000	14741.7	0.380	20.449	20.574	-0.698	11.496	11.096	20.514
4.000	11746.8	-0.110	22.838	22.877	-1.062	14.530	13.935	22.742
5.000	9797.9	-0.020	25.048	24.985	-1.161	17.228	16.431	24.792
7.500	6965.9	-0.005	29.946	29.734	-1.227	23.002	21.761	29.366
10.000	5424.9	-0.152	33.874	33.819	-1.392	27.559	26.216	33.326
12.500	4452.1	0.551	37.438	37.664	-0.881	31.581	30.230	37.057
Taille 40								
0.100	219056.6	-0.026	8.150	8.235	-0.008	0.069	0.070	8.312
0.250	97797.7	-0.004	8.364	8.416	-0.036	0.345	0.346	8.493
0.500	56033.2	0.163	8.755	8.864	-0.076	1.002	1.009	8.935
1.000	33383.2	-0.129	9.980	9.895	-0.237	2.567	2.509	9.954
1.500	24821.9	-0.017	10.981	10.927	-0.291	4.029	3.945	10.967
2.000	20066.7	0.004	11.797	11.905	-0.333	5.306	5.264	11.925
3.000	14741.7	-0.162	13.671	13.659	-0.504	7.732	7.571	13.651
4.000	11746.8	0.119	15.413	15.250	-0.398	9.860	9.553	15.213
5.000	9797.9	-0.265	16.683	16.656	-0.676	11.562	11.300	16.601
7.500	6965.9	0.113	19.731	19.855	-0.439	15.311	15.024	19.743
10.000	5424.9	0.167	22.595	22.569	-0.460	18.554	18.099	22.424
12.500	4452.1	0.222	24.930	25.044	-0.399	21.198	20.823	24.864



LA PRÉCISION DES ESTIMATIONS DE L'INÉGALITE DES REVENUS DANS LES ENQUÊTES AUPRÈS DES MÉNAGES

Jérome Accardo, Madior Fall

Introduction

On illustre ici l'intérêt pratique de l'indicateur d'inégalité de Gini et du calcul de sa précision dans le cadre d'une analyse de l'évolution de la distribution des revenus des ménages, en France, sur une période qui couvre approximativement la seconde moitié des années 80.

L'approche suivie est fondée avant tout sur l'utilisation comparative de plusieurs enquêtes de l'Insee auprès des ménages réalisées durant cette période.

- une première partie descriptive expose les différentes évaluations de l'inégalité des revenus. Retenant l'indice de Gini comme indicateur privilégié, elle montre la nécessité, pour établir valablement un diagnostic sur son évolution, de déterminer la précision avec laquelle cet indicateur est estimé.

- une deuxième partie met en oeuvre une formule de linéarisation proposée par Deville (1995) pour évaluer la variance de sondage du Gini du revenu dans les enquêtes-ménages de l'Insee.

- la troisième partie aborde le problème de l'absence de robustesse du Gini aux erreurs de mesure. Elle suggère que, malgré les apparences, ce problème est vraisemblablement de peu de poids pratique.

I. L'inégalité des revenus[1] dans les enquêtes ménages de l'Insee, 1984 -1993

On dispose de six familles d'enquête qui saisissent le revenu des ménages et, pour chacune, de deux réalisations, la première au milieu de la décennie 80 (1984, 1986), la seconde de 5 à 8 ans postérieure, soit :

- Budget de Famille (BDF) 1984 et 1989,
- Revenus Fiscaux (RF) 1984 et 1990,
- Logement (LOG) 1984 et 1992,
- Enquête Trimestrielle de Conjoncture (CJ) mars 1984 et mars 1990,
- Actifs Financiers (AF) 1986 et 1992,
- Situations Défavorisées (SD) 1986 et 1993

Le rapprochement de divers indicateurs d'inégalité calculés sur ces enquêtes met en évidence la difficulté à établir un diagnostic clair sur l'évolution de l'inégalité des revenus sur la période.

Comme le montre le tableau ci-après,

- d'une part les niveaux des indicateurs ne coïncident pas toujours, même pour des enquêtes de même millésime, et il paraît souhaitables de rationaliser ces écarts.
- d'autre part le sens d'évolution de l'inégalité dépend
- i) du type d'indicateur choisi
- ii) de l'enquête retenue

1. Il s'agit, dans tout ce qui suit et sauf mention explicite, du revenu moyen par unité de consommation d'Oxford.

Tableau 1 : Indicateurs d'inégalité, revenu/uc

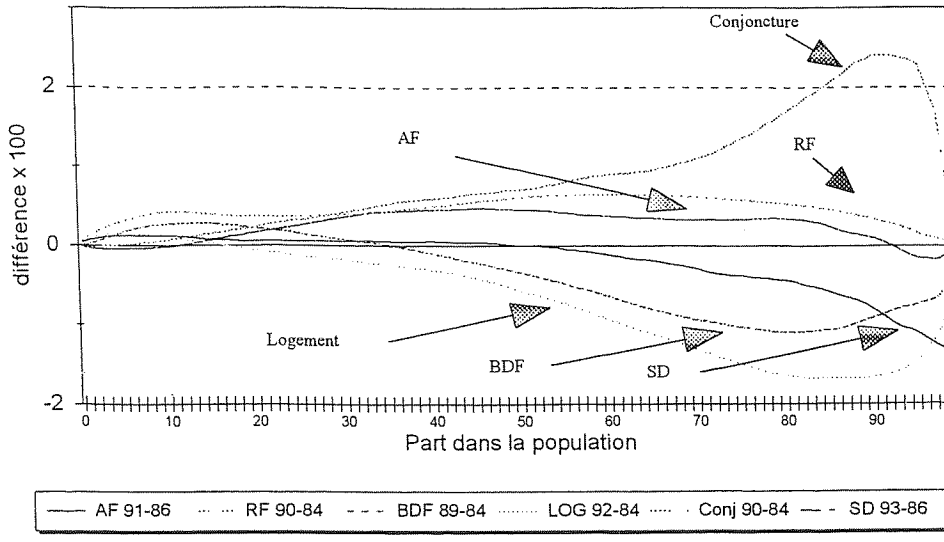
Enquêtes	D9/D1	Theil	Gini
AF 1986	4.91	0.218	0.352
AF 1992	4.78	0.214	0.347
BDF 1984	5.16	0.200	0.334
BDF 1989	4.84	0.214	0.341
RF 1984	4.15	0.235	0.341
RF 1990	3.85	0.219	0.329
LOG 1984	4.20	0.204	0.328
LOG 1992	4.39	0.227	0.342
SD 1986	4.34	0.189	0.328
SD 1993	4.08	0.213	0.332
CJ 1984	4.60	0.217	0.349
CJ 1990	4.56	0.186	0.332

L'analyse des courbes de Lorenz sur les deux vagues d'enquêtes successives confirme la complexité des évolutions *(graphique 1).*

Remarque : quelle que soit l'enquête considérée, la courbes de Lorenz de la première vague est indiscernable de la seconde (quand elles sont représentées aux formats usuels). Aussi on a choisi de représenter, pour chaque enquête, la différence des courbes de Lorenz de chacune des deux réalisations.

Sur ce graphique, on remarque que pour plusieurs enquêtes la différence des deux courbes change au moins une fois de signe (sauf pour RF où l'inégalité paraît baisser légèrement). Il paraît ainsi difficile de se prononcer, sur une base descriptive, sur le sens d'évolution de l'inégalité des revenus sur la période. Ce constat souligne l'intérêt d'une évaluation de la précision des estimateurs des indices d'inégalité fournis par les enquêtes auprès des ménages. On se borne, dans ce qui suit, au cas du Gini.

Graphique 1 : **Différence de courbes de lorenz**
revenu par uc



II. Estimation de la précision de la mesure d'enquête du Gini.

On expose ici succinctement les procédures de calcul de la précision pour l'écart-type de sondage.

Le calcul de variance de sondage utilise la formule de linéarisation proposée par J.C. Deville avec les étapes suivantes (les calculs sont effectués sur les enquêtes LOG 1992 et SD 1993)

I. L'estimation pratique de la précision du coefficient de Gini dans les enquêtes "Logement 1992" et "Situations Défavorisées 1993".

1) Le calcul de la précision repose en premier lieu sur une linéarisation du coefficient de Gini due à J.C. Deville (Deville, 1995)[2].

2 Deville, Jean-Claude (1995),"Estimation de la variance du coefficient de Gini mesuré par sondage", Document de Travail n° F9510, Insee.

On suppose ici qu'on dispose de n observations du revenu, (y_i), i=1,...,n , tirées selon un certain plan de sondage. Soit (w_i), i=1,...,n les poids de sondage associés (i.e. l'inverse des probabilités de tirage). La procédure proposée par Deville consiste à :

- a) ordonner les observations dans le sens croissant. Soit (z_i), i=1,...,n la séquence ainsi obtenue.

- b) calculer les cumuls $W_k = \sum_{i=1}^{k} w_i$ et $Z_k = \sum_{i=1}^{k} w_i z_i$

- c) calculer les totaux N= (la population totale) et Z (la masse totale des revenus).

- d) calculer R = S/Z , où $S = \sum_{i=1}^{n} z_i w_i \frac{W_i}{N}$. On en déduit la valeur du coefficient de Gini par G=2R-1.

- e) calculer les n valeurs $u_k = F_k\left(z_k - Z_k / W_k\right) - Rz_k$, où $F_k = \left(W_k - 0.5w_k\right)/N$, autrement dit F_k est la fonction de répartition empirique des (z_k), légèrement lissée.

Il suffit alors de calculer la variance V du total $\sum_{k=1}^{n} w_k u_k$. L'écart-type du coefficient de Gini est alors

$$\sigma = 2\frac{\sqrt{V}}{Z}$$

2) On est ainsi ramené au calcul de la variance du total (pondéré) d'une variable artificielle u_K Les enquêtes considérées sont tirées selon un plan de sondage complexe : stratification, tirage à plusieurs degrés, en plusieurs phases. La pondération de ces enquêtes est en outre redressée par Calmar. Pour pouvoir mener à bien le calcul de variance, on est ainsi amené à faire les simplifications suivantes :

A) Dans les strates de tirage appartenant aux strates de gestion 0, 1, 2 (i.e. respectivement, les communes rurales au RP 1990, les communes appartenant à des unités urbaines de moins de 20 000 habitants au RP 1990, les communes appartenant à des unités urbaines ayant entre 20 000 et 100 000 habitants au RP 1990).

Dans chacune de ces strates on suppose que le premier degré de tirage (tirage des Unités Primaires de l'échantillon-maître) s'est fait avec remise. On peut alors facilement calculer la variance (voir P. ARDILLY, "Les techniques de sondage",

ed.Technip, 1994, p 131 et sq.) , dans chaque strate d'un total . Pour chaque UP d'une part on estime le total de la variable revenu y dans l'UP n° j par

$$Y_j = \sum_{i \in UP_j} w_i y_i \ .$$

(Ce calcul suppose évidemment qu'on connaît, pour chaque observation, l'UP dont elle provient). D'autre part on calcule Q, la probabilité d'une UP (de la strate) d'être tirée (à chacun des tirages du tirage avec remise), soit Q = (Nombre de logements dans l'UP)/(nombre de logements dans la strate).

Soit alors

$$\overline{Y} = \frac{1}{m} \sum_{j=1}^{m} \frac{Y_j}{Q_j} ,$$

où m est le nombre d'UP tirées (dans la strate considérée). La variance (intra-strate) de Y est alors donnée par

$$V(Y) = \frac{1}{m(m-1)} \sum_{j=1}^{m} (\frac{Y_j}{Q_j} - \overline{Y})^2$$

Il suffit alors de sommer ces variances sur l'ensemble des strates considérées.

B) Dans les strates de gestion 3 et 4 (i.e. respectivement les communes appartenant à des unités urbaines de plus de 100 000 habitants au RP 1990, sauf l'unité urbaine de Paris, et l'unité urbaine de Paris) : on considère ici que, dans chacune de ces deux strates, le sondage est un sondage aléatoire simple, chaque observations étant tirée avec la probabilité : n/N, où n est le nombre d'observations dans la strate et N le nombre total de logements dans la strate.

Pour chacune des deux strates, la variance est alors classiquement donnée par

$$V(Y) \approx \frac{1-f}{f^2} \Sigma (y_k - \overline{y})^2$$

où f est le taux de sondage dans la strate.

On obtient ainsi (en francs courants) :

Logement 1992 : écart-type du revenu moyen=679 F , écart-type du Gini=0.0058

Situ.Défav1993 : écart-type du revenu moyen=772 F, écart-type du Gini=0.0074

3) Si on admet que les ordres de grandeur obtenus sont généralisables à toutes les enquêtes-ménages analysées dans ce travail, il peut être intéressant de déterminer l'analogue du design effect (ou effet de sondage). Posons

$$DEFF = \frac{\sigma_{sondage}}{\sigma_{1passe}}$$

où le numérateur est l'écart-type calculé comme ci-dessus, et où le dénominateur désigne l'écart-type obtenu en supposant que les observations ont été tirées en une passe, selon l'inverse de leur pondération effective (i.e. celle, éventuellement issue de Calmar, qui apparaît dans le fichier d'enquête final). Cet écart-type est facilement calculable en utilisant une formule asymptotique fournie par Rosen (1991), qui évite le calcul des probabilités d'inclusion double.On obtient alors :

Logement 1992 : DEFF du revenu moyen = 0.77 , DEFF du Gini = 0.77

Situ.Défav1993 : DEFF du revenu moyen = 0.50, DEFF du Gini = 0.50

4) Conclusion : on retiendra 0.007 comme écart-type du Gini dans les différentes enquêtes ménages. Un rapide calcul de linéarisation à partir de l'écart-type du revenu moyen fournit, pour l'indice de Theil, un écart-type de 0.015 environ.

Remarque : On a procédé à une simulation numérique pour vérifier la solidité de l'approche par la linéarisation. Dans le cadre certes limité de cette simulation, les résultats apparaissent tout à fait satisfaisants. En outre l'approche par linéarisation présente des performances supérieures à une approche de type bootstrap.

1) Simulations

On a constitué une population artificielle de ménages en empilant les différents fichiers d'enquête analysés ici (RF, AF, LOG, ...) ; on obtient ainsi une population d'environ 200 000 ménages. Chaque ménage est affecté de la probabilité d'inclusion égale à l'inverse de son poids de sondage tel qu'il apparaît dans l'enquête. Les probabilités d'inclusion présentent donc une dispersion considérable (cf. Document de travail n° F 9602, Insee, section I).

On tire alors, d'abord en utilisant un tirage uniforme, pour différents taux de sondage, 50 échantillons. pour chacun de ces échantillons, k=1,...,50 on calcule

a) le coefficient de Gini, G(k)

b) la variance varD(k) de ce coefficient, telle qu'on la calcule par la formule de linéarisation de Deville. On pose alors VD=moyenne des varD(k), k=1,...,50. Soit $\sigma_d=\sqrt{VD}$ l'écart-type correspondant.

c) la variance empirique VE de l'échantillon des G(k) et $\sigma_e=\sqrt{VE}$.

On effectue ensuite ces calculs pour des tirages selon les probabilités d'inclusion du fichier, en faisant varier le taux de sondage.

Tableau 2 : vérification du calcul de variance par linéarisation

Taille de l'échantillon	Type de pondération	Écart type empirique (σ_e)	Écart type Deville (σ_d)
102	uniforme	0.13	0.93
1 017	uniforme	0.04	0.05
4 067	uniforme	0.02	0.02
2 180	inégale	0.09	0.08
4 356	inégale	0.06	0.06
8 711	inégale	0.04	0.04

2) Bootstrap

La lourdeur des calculs impliqués par le bootstrap impose de se borner à des sous-populations de taille réduite. On a ainsi retenu des tranches d'âge quinquennale. Les calculs ont été menés sur LOG 1992 et sur RF 1990. Bien entendu il n'est ici pas tenu compte de la stratification : on suppose ici que le tirage est "en une passe". On constate qu'une estimation de la variance par bootstrap fournit une évaluation un peu trop "optimiste" de la vraie variance, et biaise ainsi l'inférence en faisant trop souvent considérer comme significatives les évolutions des indicateurs d'inégalité.

Tableau 3 : Comparaison du calcul de variance par bootstrap et par linéarisation

	Rev Fiscx			Logement		
âge	Gini	σ_b	σ_d	Gini	σ_b	σ_d
25-30	0.249	0.018	0.008	0.287	0.013	0.008
30-35	0.271	0.014	0.006	0.325	0.013	0.006
40-45	0.309	0.012	0.008	0.341	0.013	0.007
45-50-	0.322	0.015	0.011	0.361	0.013	0.009
65-70	0.282	0.018	0.012	0.339	0.019	0.008
70-75	0.262	0.024	0.010	0.332	0.022	0.010

Note : σ_b= écart-type calculé par bootstrap, σ_d= écart-type selon la formule de linéarisation de Deville

On déduit alors de ce calcul de précision précédent :

i) les écarts de niveau du Gini entre les enquêtes n'apparaissent pas incompatibles avec la variance de sondage

ii) on peut proposer un test de significativité des évolutions du Gini, dans chaque enquête.

Tableau 4 : p-values d'un test de significativité des évolutions de l'inégalité

	AF	RF	BDF	LOG	SD	CJ
ΔGini	- 0.005	- 0.012	+ 0.007	+0.014	+ 0.004	- 0.017
p-value	30 %	11%	24 %	8 %	34%	4 %

Note : pour chaque enquête on calcule la variation du Gini entre ses deux réalisations. On peut alors calculer, sous l'hypothèse que cette variation suit une loi normale, de moyenne nulle et de variance $2\sigma^2$, la probabilité P d'une variation supérieure ou égale en valeur absolue à l'écart observé.

Au vu de ces résultats ,il ne paraît pas possible de conclure à une variation significative de ces indicateurs : elles sont en effet contradictoires et seule CJ (dont on a pu noter précédemment le caractère atypique , cf. *Graphique 1*), atteint le seuil usuel de rejet (5%).

III. La question de la robustesse du Gini

1. La critique théorique

Le calcul de la variance de sondage du Gini clôt-il la discussion ? Malheureusement non. En effet, aux aléas de sondages viennent s'ajouter les erreurs de mesure. Or, le Gini, comme la plupart des indicateurs d'inégalité, se révèle peu robuste à certaines erreurs de mesure ; plus précisément, considérons les trois types suivants d'erreur sur la mesure du revenu : si on note x le vrai revenu du ménage et $\tilde{x}$ sa réponse telle qu'elle apparaît dans le fichier final,

(i) le ménage se trompe sur la période considérée : il déclare son revenu mensuel alors qu'on l'interroge sur son revenu annuel (i.e. $\tilde{x} = x/12$) ou l'inverse (i.e. $\tilde{x} = 12\,x$)

(ii) le ménage se trompe sur l'unité et répond en anciens francs : $\tilde{x} = 100\,x$

(iii) lors du processus de saisie, des erreurs de report se produisent qui modifient la réponse du ménage de l'ordre d'une décimale : $\tilde{x} \cong 10^n x$, avec n = -2, -1, 1, 2,...

Dans ce cas, des taux d'erreur même faibles, inférieurs à 5 % par exemple, peuvent affecter très sensiblement la valeur du Gini. Soit ainsi la distribution des revenus de l'enquête « Actifs Financiers 1992 » portant sur 9530 ménages[3]. Le Gini est de 0.37. Si on la perturbe en multipliant par f=10 le revenu déclaré d'une proportion τ=5% des ménages, le Gini passe à 0.55 (le Theil passe de 0.23 à 0.69).

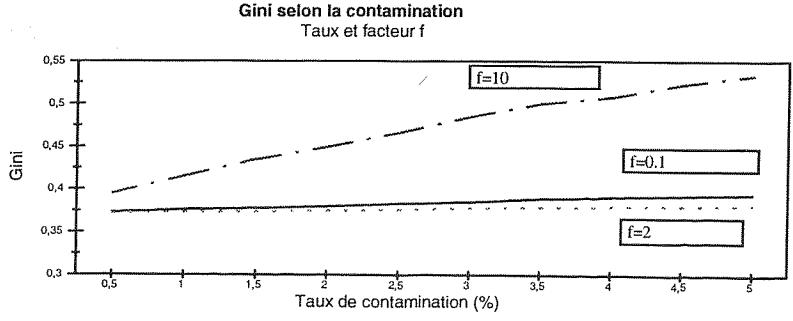
Les graphiques suivants présentent l'effet de ce type de contamination sur le Gini (pour différents taux de contamination τ et trois facteurs d'erreur f=0.1, f=2, f=10) dans deux cas :

- pour la distribution observée, de niveau de concentration 0.37 *(graphique 2-1),*

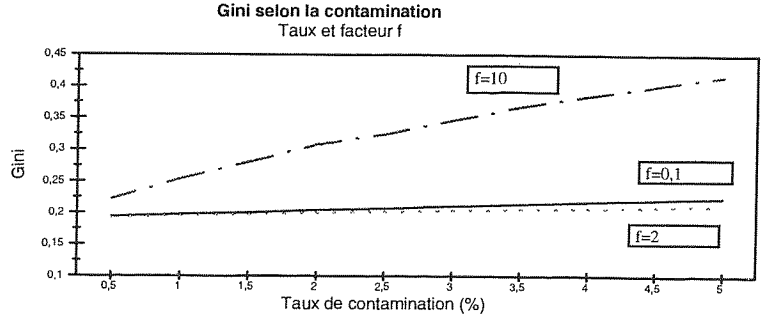
- pour la même distribution, mais transformée par l'opérateur $\sqrt{}$, de façon à obtenir une distribution moins concentrée, de gini=0.19 *(graphique 2-2).*

3 Il s'agit, ici, du revenu global du ménage.

Graphique 2.1 : (distribution observée)



Graphique 2.2 (distribution modifiée)



On peut alors constater en particulier que :

α) si le facteur d'erreur est élevé (f=10), une contamination très faible (de l'ordre de 1 %) à déjà une incidence notable sur le Gini.

β) en revanche, quand le facteur d'erreur est petit, cette incidence reste modérée, même pour des contaminations notables.

γ) il n'y a pas de relation monotone entre la valeur du facteur d'erreur et l'effet de la contamination sur le Gini (ceci étant essentiellement dû au fait que le Gini est plus sensible à ce qui se passe dans le haut de la distribution).

δ) les effets de la contamination apparaissent dépendre peu de la concentration initiale de la distribution.

Pour plus de détail sur ce type d'approche, voir Cowell, Victoria-Feser (1996)[4].

De façon très similaire, on peut arguer que les fractiles extrêmes (par exemple les 1 % des ménages les plus pauvres et les 1 % des ménages les plus riches) de la distribution des revenus sont très mal appréhendés : la mauvaise perception (et donc mauvaise déclaration) des revenus socialisés (y.c. les remboursements maladie) est particulièrement gênante chez les plus pauvres dans la mesure où, chez ces ménages, ces revenus constituent une part relativement plus importante des ressources, chez les plus riches les revenus du patrimoine sont mal mesurés, les bénéfices sont souvent difficilement isolables du chiffre d'affaires, quand il s'agit d'indépendants, etc.

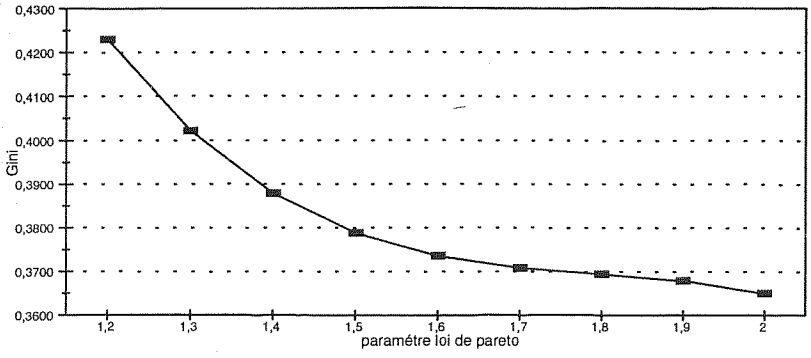
On peut ainsi, par exemple dans le cas du haut de la distribution, se demander quel est l'effet sur le Gini de l'incertitude sur la forme exacte de la distribution dans cette zone des plus hauts revenus. Par exemple en expérimentant l'incidence de différentes hypothèses quant à cette forme. On constate que le Gini, là encore, varie sensiblement selon la forme retenue.

Soit ainsi, dans l'enquête « Actifs Financiers 1992 », le dernier centile des revenus : il est composé d'environ 200 000 ménages qui disposent chacun de plus de 352 000 F (par uc). Le plus haut revenu (par uc) mesuré est : 1 177 000.

On peut se défier de la distribution empirique des revenus dans ce dernier centile, et préférer supposer d'autres distributions ; à titre d'exemple, on a supposé que cette distribution était en réalité une loi de Pareto, de seuil 352 000 F, et de paramètre α, α étant choisi pour que la moyenne des revenus dans le dernier centile soit compatible avec l'ordre de grandeur de la moyenne empiriquement observée dans le dernier centile. On a ainsi fait varier α entre 1.2 et 2. Pour un α donné, on a tiré 10 loi de Pareto, obtenant ainsi 10 distributions alternatives des revenus dans le dernier centile et donc, 10 distributions globales des revenus qui ne diffèrent que par leur queue. On a calculé alors la moyenne des 10 Gini correspondants. Le *graphique 3* montre que pour des valeurs de α proches de 1.2, les revenus apparaissent plus concentrés que ce que mesure l'enquête. L'effet est cependant sensiblement moins fort que celui, vu ci-dessus, des erreurs de mesure.

4 Cowell, Frank A. and Maria-Pia Victoria-Feser (1996), « Robustness Properties of Inequality Measures », *Econometrica*, Vol. 64, n° 1, 77-101

Graphique 3 : Simulation Gini
Dernier centile suit la loi de Pareto



Des analyses de ce genre paraissent remettre en cause l'utilisation habituelle du Gini, même suppléée par le calcul de la variance de sondage : elles impliquent en effet que l'estimation usuelle est extrêmement imprécise, ce qui en particulier interdit pratiquement tout diagnostic quant à l'accroissement ou la baisse des inégalités entre deux dates. Dans cette optique la variance de sondage ne représenterait qu'une borne (très) inférieure de l'erreur probable, sans intérêt pratique.

2. *Le constat empirique*

Pourtant, de façon surprenante étant donné ce qui précède, les estimations du Gini de la distribution des revenus effectuées dans les diverses enquêtes auprès des ménages, en France, ne présentent pas une dispersion particulièrement frappante.

Tableau 5 : Évolutions du Gini entre différentes enquêtes

	Actifs Fin.	Rev. Fiscx	Budg. Famil.	Logement	Sit. défav.	Enq. Conj.	Ecart Max.
Gini - 1	0.352	0.341	0.334	0.328	0.328	0.349	0.024
Gini - 2	0.347	0.329	0.341	0.342	0.332	0.332	0.018
Évolution	+ 0.005	- 0.012	+ 0.007	+ 0.014	+ 0.004	- 0.017	

La dernière colonne (Ecart Max) donne, pour chaque vague, la différence maximale observée entre les Gini d'une même période. Il est clair qu'on ne retrouve pas, ici, la très forte variabilité suggérée par l'analyse de (manque de) robustesse précédente. A chaque période, l'écart observé est au plus de 0.024 (c'est aussi l'écart maximum entre toutes les enquêtes du tableau).

A l'inverse, sous l'hypothèse de constance de la distribution réelle des revenus sur la période couverte par les enquêtes d'une même vague, l'écart observé est compatible avec l'intervalle de confiance à 95 % prédit par la variance de sondage, [g-2σ, g+2σ], avec σ=0.007.

En toute rigueur, ce constat est insuffisant pour invalider les conclusions de l'analyse de robustesse : on peut en effet soutenir que l'apparente convergence des Gini résulte de ce que toutes les enquêtes sont affectées de façon comparable par les erreurs de mesure. Le niveau serait donc grossièrement erroné. En outre, les fortes variations dues, en principe, aux erreurs de mesure interdiraient de tabler sur une constance de leurs effets, donc interdiraient une utilisation du Gini (en tous cas de ses estimateurs empiriques usuels) dans une analyse en évolution. Voir Cowell,Victoria-Feser, op.cit.

Pourtant, encore une fois, le rapprochement de différentes enquêtes suggère des conclusions nettement plus optimistes. Les enquêtes ne sont en effet pas également exposées aux trois types (i) (ii), (iii) d'erreur de mesure mentionnés plus haut. Si, par exemple, une erreur de type (iii) (où une erreur de saisie a divisé ou au contraire multiplié par 10 le revenu déclaré par le ménage) est plausible dans une enquête comme Budget de Famille ou Revenus Fiscaux, elle est tout à fait improbable dans Actifs Financiers où le ménage est invité a déclarer un revenu en tranche. De même une erreur de type (i) (le ménage déclare un revenu annuel à une question sur son revenu mensuel ou vice et versa) est possible dans Actifs Financiers, déjà moins probable dans Budget de Famille, enquête dont le questionnaire sur le revenu est très détaillé et réduit très certainement ce genre de confusion. Elle est très peu vraisemblable dans le cas de Revenus Fiscaux, où l'on peut raisonnablement penser que les ménages font particulièrement attention à ne pas multiplier fortuitement par 10 leur revenu fiscal ![5]

On peut ainsi proposer le tableau suivant qui indique, de façon qualitative, la probabilité a priori des différents types d'erreurs sur les six familles d'enquêtes citées supra.

5 Il est possible certes qu'ils fraudent...mais, dans ce cas, on observe $\tilde{x} << x$, et on a vu que l'effet de ce type d'erreur était relativement négligeable.

Tableau 6 : Types d'erreur plausibles dans les 6 enquêtes.

	Actifs Fin.	Rev. Fiscx	Budg. Famil.	Logement	Sit. Défav.	Enq. Conj.
erreur (i)	+	-	?	+	?	+
erreur (ii)	?	-	?	+	?	?
erreur (iii)	-	+	+	+	+	-

Légende :

« + » indique que l'erreur est a priori plausible, dans une proportion non négligeable (de l'ordre de quelques %),
« ? » indique qu'il est difficile de se prononcer a priori sur la plausibilité de l'occurrence de ce type d'erreur,
« - » indique que ce type d'erreur est a priori très peu probable.

On voit ainsi apparaître deux cas polaires :

- l'enquête Revenus Fiscaux, la moins susceptible d'être entachée d'erreur de mesure

- l'enquête Logement, dans laquelle le statut relativement mineur de la partie « revenus » du questionnaire (elle n'a pour fonction essentielle que de fournir une indication de la capacité du ménage à investir dans un logement) suggère a priori une plus forte proportion d'erreur de différents types.

Si on admet alors que les 3 types d'erreur sont équiprobables, de taux α, Logement 1992 contient vraisemblablement une proportion 3α d'erreurs et Revenus Fiscaux seulement α. L'écart de Gini observé est : $\Delta=0.013$ qui doit ainsi provenir du différentiel de contamination 2α. D'après les graphiques 2-1 et 2-2, indiquant le niveau du Gini en fonction du taux de contamination α, on a au plus $\alpha \approx 0.3$ %, autrement dit une contamination globale (de facteur 10 et plus) de très faible ampleur, dont l'effet sur le niveau du Gini est pratiquement négligeable.

Session 4

Les statistiques locales

ESTIMATION SUR DES PETITS DOMAINES
- APPLICATION A L'ENQUÊTE ÉDUCATION 92 -

Sophie Destandau

Sommaire

Introduction

Les sources statistiques

Pendant longtemps, les seules **sources statistiques** disponibles au niveau national et local furent les **recensements** et les **fichiers administratifs.**

Or, la tendance actuelle est de réaliser des recensements tous les 10 ans; et les champs des fichiers sont fréquemment modifiés en fonction de la politique adoptée dans l'administration associée.

Dans les années 40/50, un nouveau type de méthode de collecte d'information est apparu et s'est rapidement développé pour pallier les inconvénients des deux autres : c'est **l'enquête par sondage.**

Une demande de données locales de plus en plus pressante...

Parallèlement au développement des méthodes d'estimation par sondage, est apparue une nouvelle catégorie de demande : **celle de données locales .**

En effet, ces dernières sont nécessaires à l'élaboration des politiques et des programmes gouvernementaux afin de distribuer des fonds ou pour mettre en place une planification régionale.

Les méthodes par sondage pour obtenir des données localisées

Pour obtenir des résultats sur des domaines particuliers, 2 méthodes sont envisageables utilisant aussi bien l'une que l'autre la technique des enquêtes par sondage :

1 - On peut réaliser **une enquête locale.**

Il faut donc

* une base de sondage locale
* une possibilité de stratification
* des informations globales sur la population à étudier
* et surtout une taille d'échantillon équivalente à celle associée à une enquête nationale ; elle ne dépend pas de la taille du domaine

2 - On peut aussi **utiliser de manière locale des enquêtes globales** ce qui constitue l'objet de mon papier.

Les inconvénients d'une enquête nationale par sondage

a) une enquête par sondage globale a pour but la production d'estimateurs fiables sur des variables d'intérêt générales **à des niveaux d'agrégation supérieurs tel le niveau national.**

b) dans la phase d'échantillonnage, **les zones** sur lesquels au cours d'une exploitation ultérieure, il s'avérerait intéressant d'avoir des données, **ne sont pas initialement prévues** ; la zone en question est **donc** généralement **non planifiée** ('unplanned' en anglais). Car, même si le plan d'échantillonnage est stratifié, la zone intéressante peut être différente de celle définie par une strate ou un groupe de strates. Elle nécessite parfois un redécoupage complet de la population.

Qu'est ce qu'un domaine ?

Les données locales ou localisées trouvent des applications multiples et variées dans tous les secteurs (dotations...). Elles peuvent en effet, ne pas être 'locales' dans le sens géographique mais plutôt dans le sens 'réduit à une zone'. C'est pourquoi, le terme employé dans la littérature sur le sujet est **domaine**.

Types de domaine :

* **géographique** (nommé 'area' ou 'areal domain' en anglais)

Exemple : un département, une région

* **croisement** de variables socio-démographiques ou autres (nommé 'domain' ou 'characteristic domain' en anglais) :

Exemple : tranche d'âge-sexe

Qu'entend-t-on par petit domaine ?

'Like beauty, small is in the eyes of the beholder'

Martin Wilk
International Symposium (86)

Purcell et Kish (1979) dans leur article intitulé 'Postcensal Estimates for Local *Areas (or Domains)'* distinguent eux **4 types de domaines selon N_d la taille du domaine** (c'est à dire la somme des unités contenues dans l'intersection de la population avec le domaine) :

Soit $P_d = \frac{N_d}{N}$

- les grands ('major') domaines pour lesquels $P_d \geq 0{,}1$
- les moyens ('minor') domaines pour lesquels $0{,}01 \leq P_d \leq 0{,}1$
- les petits ('mini') domaines pour lesquels $0{,}0001 \leq P_d \leq 0{,}01$
- et les très petits ('rare') domaines pour lesquels $P_d \leq 0{,}0001$.

Le choix de la méthode

Platek, Rao, Särndal, Singh (International Symposium de *Statistique Canada* - 86) ont déclaré que des méthodes d'estimation différentes doivent être employées selon la taille du domaine.

*'A **small area** is any area for which direct design-based estimates cannot be reliably produced from the current sample survey program or a reasonnable expansion thereof. Small areas thus become areas that need other methods of estimation'.*

Le **choix de l' estimation** sur un domaine est lié à plusieurs facteurs comme

- la taille de l'échantillon s : n
- la taille du domaine dans l'échantillon : n_d
- le plan d'échantillonnage c'est à dire encore :
  - la base de sondage
  - le mode de tirage de l'échantillon : sondage aléatoire simple, stratification, sondage par grappes, en plusieurs phases, à plusieurs degrés...
- les variables auxiliaires disponibles

De nombreux statisticiens insistent sur l'importance du choix des variables auxiliaires. Elles doivent être fortement corrélées avec la variable d'intérêt de l'enquête afin d'améliorer la précision des estimations.

Elles peuvent provenir aussi bien de recensements, des registres que d'autres enquêtes par sondage.

♦ les variables d'intérêt à estimer et leurs fréquences (mensuelles, annuelles ...)

♦ la politique décidée par le responsable d'enquête en matière de coût et de précision des résultats

♦ les modèles ou les hypothèses choisis

♦ l'équilibrage entre biais et variance de l'estimateur.

Les différentes catégories d'estimateurs

M.P. Singh, J. Gambino, et H.J. Mantel, dans leur article intitulé *'Les Petites Régions Problèmes et Solutions' (94),* ont établi une classification des estimateurs de ce type en mesurant les conséquences au niveau du **biais** et de la **variance** inconditionnels.

⇒ les **estimateurs de plan** ou **directs**

Ils seront utilisés **lorsque**
* n_d sera grand
* ou lorsque le domaine est prévu dans le plan d'échantillonnage.
Leur biais est nul et leur variance faible.

⇒ les **estimateurs indirects ou de modèle.**

Ils seront utilisés dès lors que n_d **sera considéré comme faible.**

Beaucoup de méthodologues conseillent d'utiliser cette deuxième catégorie d'estimateurs avec beaucoup de circonspection et seulement lorsque
* tous les estimateurs de plan ont été envisagés
* les données auxiliaires ont fourni le maximum de précision supplémentaire.

D'autres classifications existent mais dans cette étude, trois types d'estimateurs de domaines seront détaillés: les estimateurs directs, et deux types d'estimateurs indirects : les estimateurs synthétiques et les estimateurs combinés...

Dans un deuxième temps, deux applications simples de ces méthodes à partir de l'enquête dite Education réalisée en mai/juin 1992 seront présentées.

Notations

1 - La population U

→ Elle comprend N **individus**. Un individu sera noté k, k variant de 1 à N.

→ Elle est découpée en D **domaines** exhaustifs et disjoints. Chaque domaine sera noté d, d variant de 1 à D. Il comprend N_d individus.

$$N = \sum_{d=1}^{D} N_d$$

→ La **variable d'intérêt** relative à l'individu k est notée Y_k.

La somme, la moyenne et la variance de cette variable sur l'ensemble de la population seront notées :

$$Y = \sum_{k=1}^{N} Y_k \qquad \overline{Y} = \frac{1}{N} \sum_{k=1}^{N} Y_k$$

$$S^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \overline{Y})^2$$

tandis que celles relatives au domaine d

$$Y_d = \sum_{k=1}^{Nd} Y_k \qquad \overline{Y}_d = \frac{1}{N_d} \sum_{k=1}^{Nd} Y_k$$

$$S_d^2 = \frac{1}{N_d - 1} \sum_{k=1}^{N_d} (Y_k - \overline{Y}_d)^2$$

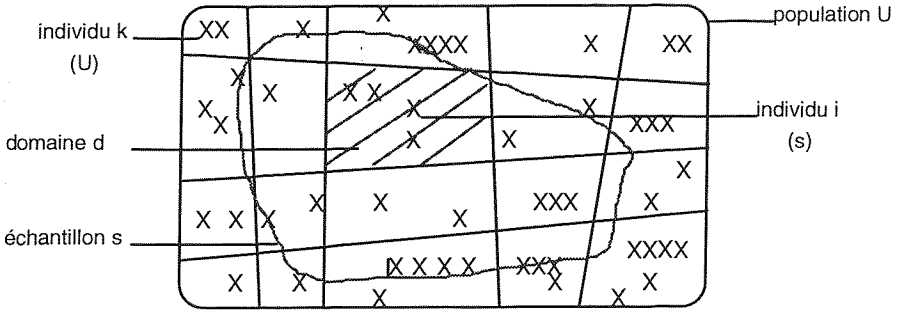
→ La population pourra être aussi subdivisée selon **un autre critère de classification** que celui des domaines ; ce sont G groupes exhaustifs et disjoints.

$$N = \sum_{g=1}^{G} N_g = \sum_{d=1}^{D} \sum_{g=1}^{G} N_{dg}$$

avec N_{dg} le nombre d'individus appartenant à la fois au domaine d et au groupe g.

2 - L'échantillon s

Sa représentation est la suivante :



→ L'échantillon s comprend n **individus**. Un individu appartenant à l'échantillon sera noté i, i variant de 1 à n.

→ Chaque **individu i** appartient aussi à un des D domaines précédemment définis. Chaque **domaine** d comprend n_d individus.

$$n = \sum_{d=1}^{D} n_d$$

→ La **probabilité** de tirage d'un individu i est notée

$$P(i \in s) = p_i$$

→ La **variable d'intérêt** relative à l'individu i est notée: Y_i

3 - Les variables auxiliaires X

→ La **corrélation** entre ces variables auxiliaires choisies et la variable d'intérêt doit être forte.

→ La variable auxiliaire peut être **disponible :**

* sur l'ensemble des individus k de la population

La notation adoptée est X_k, k variant de 1 à N

* sur des groupes g exhaustifs et disjoints différents des domaines

Les totaux sur ces groupes sont notés X_g , g variant de 1 à G

* sur les domaines d

Les totaux sur ces domaines sont notés X_d , d variant de 1 à D

L'objet de l'étude est donc d'estimer le total Y_d
(ou $\overline{Y_d} = \frac{Y_d}{N_d}$)
pour tout domaine d =1 à D

Les trois grandes catégories d'estimateurs

1 - Les estimateurs directs

Principe : les estimateurs directs n'utilisent que les données (de l'enquête ou d'une source extérieure) relatives à un seul domaine à la fois.

Leur caractéristique principale est d'être sans biais.

Pour être les plus performants au niveau biais **et** au niveau variance, ils doivent être utilisés uniquement lorsque n_d **est suffisamment grand** i.e. encore pour des domaines classés dans les 'large areas' ou 'major domains' .

1 - 1 - L'estimateur d'Horvitz Thompson

Cet estimateur est la **référence**
car il utilise seulement les valeurs de la variable d'intérêt et les poids

L'**estimateur d'Horvitz-Thompson ou estimateur par facteur d'extension** (**'expansion estimator'**) du total Y_d est la somme pondérée sur l'échantillon.

$$\hat{Y}_d(h) = \sum_{i=1}^{n_d} \frac{Y_i}{p_i}$$

En subdivisant la population en G **groupes** exhaustifs et disjoints différents des domaines, l'estimateur s'écrit :

$$\hat{Y}_d(h_g) = \sum_{g=1}^{G} \sum_{i=1}^{n_{dg}} \frac{Y_i}{p_i}$$

Biais et Variance :
Inconditionnellement, cet estimateur est sans biais

$$E(\hat{Y}_d(h)) = E(\sum_{k=1}^{Nd} \frac{Y_k}{p_k} \varepsilon_k) = \sum_{k=1}^{Nd} \frac{Y_k}{p_k} E(\varepsilon_k) = Y_d$$

$\varepsilon_k = 1$ si $k \in s_d$

$\varepsilon_k = 0$ sinon

et sa variance s'écrit :

$$V(\hat{Y}d(h)) = \frac{1}{2} \sum_{\substack{k,l=1 \\ k \neq l}}^{N_d} \sum (p_k p_l - p_{kl})(\frac{Y_k}{p_k} - \frac{Y_l}{p_l})^2$$

$$p_{kl} = P(k \in s \text{ et } l \in s \; k \neq l)$$

$$p_{kk} = p_k$$

➔ **Cas particulier d'un tirage à Probabilités Egales Sans Remise (PESR) :**

Les poids sont alors tous égaux à $\frac{n}{N}$.

$$\boxed{\hat{Y}d(d) = (N/n) \sum_{i=1}^{n_d} Y_i}$$

*** Calcul de la variance de l'estimateur direct $\hat{Y}_d(d)$**

(cf 'Model Assisted Survey Sampling' - C-E Särndal, B. Swenson, J. Wretman - p 392)

$$V(\hat{Y}_d(d)) = N^2 \frac{(1-f)}{n} \frac{(N_d - 1)S^2_{Yd} + N_d(1 - \frac{N_d}{N})\overline{Y}_d^2}{N-1}$$

$$\text{avec } S^2_{Yd} = \frac{1}{N_d - 1} \sum_{k=1}^{N_d} (Y_k - \overline{Y}_d)^2$$

*** Variance estimée de cet estimateur**

$$\hat{V}(\hat{Y}_d(d)) = N^2 \frac{(1 - \frac{n}{N})(n_d - 1)s^2_{yd} + n_d(1 - \frac{n_d}{n})\overline{y}_d^2}{n-1}$$

$$\text{avec } s^2_{yd} = \frac{1}{n_d - 1} \sum_{i=1}^{n_d} (y_i - \overline{y}_d)^2$$

$$\overline{y}_d = \frac{1}{n_d} \sum_{i=1}^{n_d} y_i$$

1 - 2 - L'estimateur de post-stratification

Il utilise **en plus la taille du domaine d dans la population :** N_d

L'estimateur de stratification a posteriori ('post-stratified estimator') s'écrit :

$$\hat{Y}_{d(pst)} = N_d \times \frac{\sum_{i \in sd} \frac{Y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}} = N_d \frac{\hat{Y}_d(h)}{\hat{N}_d(h)}$$

Inconditionnellement, cet estimateur est sans biais.

En subdivisant la population en **groupes** et si on connait N_{dg} , l'estimateur s'écrit :

$$\hat{Y}_{d(pst_g)} = \sum_{g=1}^{G} N_{dg} \frac{\sum_{i \in sdg} \frac{Y_i}{p_i}}{\sum_{i \in sdg} \frac{1}{p_i}} = \sum_{g=1}^{G} N_{dg} \frac{\hat{Y}_{dg}(h)}{\hat{N}_{dg}(h)}$$

Ces 2 estimateurs sont plus stables que les estimateurs d'Horvitz-Thompson car la connaissance de N_d (ou de N_{dg}) permet un gain de précision.

1 - 3 - L'estimateur par le ratio

Il utilise **les valeurs d'une variable auxiliaire X** bien corrélée avec la variable d'intérêt Y .

L'estimateur par le ratio ('ratio estimator') ressemble à l'estimateur de post-stratification : $\hat{Y}_{d(pst)}$ avec cette différence près qu'on utilise l'information contenue dans la variable auxiliaire X_d au lieu de l'effectif N_d . Il s'écrit :

$$\hat{Y}_d(r) = X_d \times \hat{R}_d$$

avec $\hat{R}_d$ estimation de $R_d = \frac{Y_d}{X_d}$

En subdivisant la population en **groupes** et à condition de connaître X_{dg} l'estimateur par le ratio s'écrit :

$$\hat{Y}_d\,(r_g) = \sum_{g=1}^{G} X_{dg} \times \hat{R}_{dg}$$

$$\text{avec } \hat{R}_{dg} \text{ estimation de } \frac{Y_{dg}}{X_{dg}}$$

Des estimations du ratio : $\hat{R}_d = \frac{\hat{Y}_d(h)}{\hat{X}_d(h)} = \hat{R}_d(h)$ est une des estimations de $R_d = \frac{Y_d}{X_d}$.

Mais cette estimation peut être le rapport d'une estimation directe de Y sur la vraie valeur de X sur le domaine si on en dispose

$$\hat{R}_d(a) = \frac{\hat{Y}_d}{X_d}$$

D'autres estimateurs directs existent comme **l'estimateur de régression** mais ils ne sont pas développés ici compte tenu des estimateurs utilisés dans les deux applications.

Etant donnée la **forte variabilité de la variance de ces types d'estimateurs,** les recherches se sont orientées vers sa diminution parfois au détriment du biais.

C'est ainsi que se sont développées de nouvelles techniques d'estimation fondées sur le principe d'"**emprunter de l'information**" aux autres domaines que ceux sur lesquels se réalise l'estimation en émettant des hypothèses ou en supposant des modèles. Le principe est le suivant :

'to borrow strength from related or similar small areas through explicit or implicit models that connect the small areas via supplementary data'.

Gonzalez (73)

2 - Les estimateurs synthétiques

Le qualificatif '**synthétique**' a pour origine l'hypothèse selon laquelle le petit domaine ressemble souvent d'une certaine manière à un autre domaine plus grand dans lequel il est contenu.

Gonzalez (73) qui est à l'origine de la création de ce type d'estimateurs en a donné la définition suivante :

"An unbiaised estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimes for subareas under ***the assumption that the small areas have the same characteristics as the larger area****, we identify these estimates as synthetic estimates."*

2 - 1 - L'estimateur synthétique de moyenne

L'hypothèse la plus simple est de supposer que la moyenne sur un petit domaine est égale à la moyenne globale soit $\overline{Y}_d = \overline{Y}$

L'estimateur synthétique de moyenne ('mean-synthetic estimator') se déduit de l'estimateur direct de post-stratification que l'on peut écrire aussi sous la forme : $\hat{Y}_d(pst) = N_d \times \overline{y}_d$

$$\hat{Y}_d(syn_m) = N_d \times \overline{y}$$

$$\text{avec } \overline{y} = \frac{\sum_{i \in s} \frac{Y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}}$$

Lorsque la population est subdivisée en **groupes**, l'hypothèse sous-jacente est $\overline{Y}_{d,g} = \overline{Y}_g$ avec g=1 à G, l'estimateur devient :

$$\hat{Y}_d(syn_m^g) = \sum_{g=1}^{G} N_{d,g} \times \frac{\sum_{i \in s_g} \frac{Y_i}{p_i}}{\sum_{i \in s_g} \frac{1}{p_i}}$$

Calcul du biais :

$$\text{Biais}(\hat{Y}_d(\text{syn}_m)) = E(N_d \times \frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{i=1}^{n} \frac{1}{p_i}}) - Y_d = N_d(\overline{Y} - \overline{Y}_d)$$

car l'estimateur sans biais de la moyenne sur l'ensemble des domaines est la moyenne sur l'échantillon $E(\frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{i=1}^{n} \frac{1}{p_i}}) = \overline{Y}$ et N_d connu.

Sous l'hypothèse $\overline{Y}_d = \overline{Y}$, l'estimateur est **sans biais.**

Application : Dans le cas d'un Sondage Aléatoire Simple, l'estimateur direct de stratification a posteriori a une variance en 1/nd tandis que l'estimateur synthétique associé a une variance en 1/n. Par conséquent, si l'hypothèse est vérifiée, l'estimateur synthétique sera préférable à l'estimateur direct.

2 - 2 - L'estimateur synthétique par le ratio

L'hypothèse ici est la suivante : $R_d = R$,
i.e. encore que le ratio ne dépend pas du domaine

L'estimateur synthétique par le ratio ('ratio-synthetic estimator') se déduit lui aussi d'un estimateur direct, celui par le ratio :

$$\hat{Y}_d(\text{syn}_r) = X_d \hat{R}$$
avec $\hat{R}$ estimation de $R = \frac{Y}{X}$

Lorsque la population est subdivisée en **groupes**, l'hypothèse sous-jacente est $R_{d,g} = R_g$ avec g=1 à G, l'estimateur devient :

$$\hat{Y}_d(\text{syn}_r^g) = \sum_{g=1}^{G} X_{dg} \times \hat{R}_g$$

Des estimations du ratio : ici aussi l'estimation de R ou de R_g peut se faire de multiples façons. Citons entre autres $\hat{R} = \frac{\hat{Y}(h)}{\hat{X}(h)} = \hat{R}(h)$. **Singh et Tessier (76)** ont eux utilisé l'information contenue dans la variable auxiliaire X pour estimer R par

$$\tilde{R} = \frac{\sum_{i=1}^{n} \frac{Y_i}{p_i}}{\sum_{\alpha=1}^{N} X_\alpha} = \frac{\hat{Y}(h)}{X}$$

Calcul du biais dans le cas où l'estimation du ratio est celle d'Horvitz-Thompson :

$$\text{Biais}(\hat{Y}_d(\text{syn}_r)) = E(X_d \times \frac{\hat{Y}(h)}{\hat{X}(h)}) - Y_d = X_d(\frac{Y}{X} - \frac{Y_d}{X_d}) = X_d(R - R_d)$$

Si l'hypothèse est vérifiée, le biais de cet estimateur est nul.

Parmi les autres estimateurs synthétiques, on peut aussi citer **l'estimateur synthétique de régression** ou bien les **différents estimateurs ajustés** ou enfin **le Structure PREserving Estimator (SPREE)** de Purcell et Kish (80).

Ainsi, **sous des hypothèses bien précises**, ces estimateurs se révèlent être plus performants pour des petits domaines que les estimateurs directs car ils sont aussi **sans biais** compte tenu des hypothèses et leurs variances sont nettement plus faibles.

Mais, **dès lors que les hypothèses ne sont plus vérifiées**, il faut chercher ailleurs la solution.

3 - Les estimateurs combinés

Dans l'article 'Issues and options in the provision of small area data' paru dans le recueil de la conférence internationale de 1992 à Varsovie dont le thème était 'Small Area Statistics and Survey Design', **Singh, Gambino, Mantel** (93) définissent de manière générale **un estimateur combiné** ('combined estimator') comme étant **la moyenne pondérée de 2 estimateurs quelconques** $\hat{Y}_d(1)$ **et** $\hat{Y}_d(2)$

$$\hat{Y}_d(\text{com}) = a_d \hat{Y}_d(1) + (1 - a_d)\hat{Y}_d(2)$$

avec a_d poids choisi judicieusement

3 - 1 - Le choix de la combinaison des estimateurs

Ici, nous nous intéresserons uniquement à la **combinaison d'un estimateur direct et d'un estimateur synthétique**. Ainsi, l'estimateur combiné *'vise à faire l'équilibre entre la possibilité pour l'estimateur synthétique d'être biaisé (lorsque les hypothèses ne sont pas vérifiées) et l'instabilité de l'estimateur direct'.*

$$\hat{Y}_d(\text{com}) = a_d \hat{Y}_d(\text{dir}) + (1 - a_d)\hat{Y}_d(\text{syn})$$

Rao et Choudhry dans *'Small area estimation : overview and empirical study'* (93) tout comme **Ghosh et Rao** dans *'Small Area Estimation : an appraisal' (94)* nomment ces combinaisons '**composite estimators**'.

3 - 2 - Le choix des poids de la combinaison

1 - des poids fixés à l'avance arbitrairement

Cette méthode simple est trop rigide ; elle ne permet pas de tirer profit de la justesse d'un estimateur direct pour des domaines relativement fournis en données.

2 - des poids dépendant de la taille de l'échantillon : 'sample-size dependent estimators'

L'idée consiste à utiliser :

- à fond l'estimateur direct pour un domaine particulièrement bien représenté dans l'échantillon et
- de profiter des avantages de l'estimateur synthétique dans les autres cas.

2 - 1 - Si on ne dispose pas d'information auxiliaire autre que Nd,

*** Le cas le plus simple** est :

$$\boxed{a_d = \frac{n_d}{N_d}}$$

Si nd est petit par rapport à Nd, alors on privilégiera l'estimation synthétique

➔ Cas particulier : l'estimateur BLUP (Best Linear Unbiaised Predictor) d'Hidiroglou

Dans l'article intitulé 'Estimation pour les petits domaines : théorie et pratique à Statistique Canada', **M.A. Hidiroglou** (92) décortique la construction de cet estimateur.

Dans un premier temps, il décompose la population du domaine Ud en 2 :

* l'intersection entre l'échantillon et le domaine : sd
* le complémentaire

$$Y_d = \sum_{i \in s_d} y_i + \sum_{k \in U_d - s_d} Y_k$$

et il estime les 2 termes de manière indépendante.

1 - Sur l'intersection échantillon/domaine d : sd
Il se ramène à l'estimation d'une moyenne sur l'échantillon $\overline{Y}_{s_d}$

car $\sum_{i \in s_d} y_i = n_d \overline{Y}_{s_d}$

$$\hat{\overline{Y}}_{sd} = \frac{\sum_{i \in sd} \frac{y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}}$$

2 - Sur le complémentaire
L'estimation de $\sum_{k \in Ud-sd} Y_k$ se fait sur un modèle de régression en fonction de la disponibilité de variables auxiliaires ou non

On écrira le deuxième terme sous la forme d'une moyenne

$$\sum_{k \in Ud-sd} Y_k = (N_d - n_d)\overline{Y}_{U_d - s_d}$$

$\overline{Y}_{U_d - s_d}$ sera alors estimé en utilisant l'ensemble de l'échantillon (idée du synthétique)

$$\hat{\overline{Y}}_{U_d - s_d} = \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}} = \hat{\overline{Y}}(h)$$

D'où l'**estimateur combiné BLUP d'Hidiroglou** s'écrit :

$$\hat{Y}_d(com_H) = nd \frac{\sum_{i \in s_d} \frac{y_i}{p_i}}{\sum_{i \in s_d} \frac{1}{p_i}} + (Nd - nd) \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{1}{p_i}}$$

Il combine l'estimateur direct de poststratification avec l'estimateur synthétique de moyenne en choisissant le poids égal à $a_d = \frac{n_d}{N_d}$

$$\boxed{\hat{Y}_d(com_H) = \frac{n_d}{N_d} \hat{Y}_d(pst) + (1 - \frac{n_d}{N_d}) \hat{Y}_d(syn_m)}$$

*** Autre poids de combinaison**

On peut choisir

$$\boxed{a_d = \frac{\hat{N}_d(dir)}{N_d}}$$

avec $\hat{N}_d(dir)$ une estimation directe de N_d

en tenant compte du fait que $\frac{\hat{N}_d(dir)}{N_d}$ peut être supèrieur à 1.

➔ Cas particulier : l'estimateur de Drew, Singh et Choudhry (82)

Drew, Singh et Choudhry ont construit un estimateur combinant l'estimateur direct par le ratio et l'estimateur synthétique par le ratio en déterminant les poids à partir de la taille de l'échantillon ; c'est un 'sample-size dependent' estimateur.

$$\hat{Y}_d(com_{DSC}) = a_d \hat{Y}_d(r) + (1 - a_d) \hat{Y}_d(syn_r)$$

avec $a_d = 1$ si $\frac{\hat{N}_d(h)}{N_d} \geq \delta$

$$a_d = \frac{1}{\delta} \times \frac{\hat{N}_d(h)}{N_d} \quad \text{si} \quad \frac{\hat{N}_d(h)}{N_d} < \delta$$

et $\hat{N}_d(h) = \sum_{i=1}^{n_d} \frac{1}{p_i}$ estimateur direct et sans biais de N_d

Le paramètre δ étant choisi de telle manière que l'on contrôle la contribution de l'estimateur synthétique par le ratio. En pratique, il oscille entre 2/3 et 3/2.

Application : Un estimateur du même type est utilisé sur l'enquête canadienne sur la force de travail (Canadian Labour Force Survey) pour estimer annuellement les taux de chômage pour des petites régions à Statistique Canada avec δ =2/3. Pour la majorité des domaines, le poids de l'estimateur synthétique est nul ; pour les autres son poids reste faible entre 10 et 20%.

➔ **Autre cas particulier : l'estimateur de Särndal et Hidiroglou (89)**

Il possède des poids légèrement différents. En effet,

$$\hat{Y}_d(com_{SH}) = a_d \hat{Y}_d(r) + (1-a_d)\hat{Y}_d(syn_r)$$

avec $a_d = 1$ si $\dfrac{\hat{N}_d(h)}{N_d} \geq 1$

$$a_d = \left(\frac{\hat{N}_d(h)}{N_d}\right)^{t-1} \quad \text{si} \quad \frac{\hat{N}_d(h)}{N_d} < 1$$

et $\hat{N}_d(h) = \sum_{i=1}^{n_d} \frac{1}{p_i}$

Ici, c'est t qui sera choisi de telle sorte que la contribution de l'estimateur synthétique par le ratio soit contrôlée. Ils proposent t=2.

2 - 2 - Si on connait une variable auxiliaire X et N_d

*** un cas simple**

$$\boxed{a_d = \frac{\hat{X}_d(dir)}{X_d}}$$

*** l'estimateur BLUP d'Hidiroglou**

L'idée est la même que dans le cas où on ne connaît que Nd. On estime sur l'échantillon d'une part puis sur le complémentaire.

Justement, $\sum_{k \in U_d - s_d} Y_k$ est estimé par un estimateur synthétique de type par le ratio

$$(N_d \frac{\sum_{k \in U_d} X_k}{N_d} - n_d \frac{\sum_{i \in s_d} \frac{x_i}{p_i}}{\sum_{i \in s_d} \frac{1}{p_i}}) \hat{R}$$

$$\text{avec } \hat{R} = \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{x_i}{p_i}}$$

D'où l'**estimateur combiné BLUP d'Hidiroglou** s'écrit :

$$\hat{Y}_d(com_H) = n_d \frac{\sum_{i \in s_d} \frac{y_i}{p_i}}{\sum_{i \in s_d} \frac{1}{p_i}} + (N_d \overline{X}_d - n_d \frac{\sum_{i \in s_d} \frac{x_i}{p_i}}{\sum_{i \in s_d} \frac{1}{p_i}}) \frac{\sum_{i \in s} \frac{y_i}{p_i}}{\sum_{i \in s} \frac{x_i}{p_i}}$$

Ce deuxième estimateur combine l'estimateur direct par le ratio avec l'estimateur synthétique par le ratio en choisissant comme poids $\boxed{a_d = \frac{n_d \hat{X}_d(h)}{\hat{N}_d(h) X_d}}$

$$\boxed{\hat{Y}_d(com_H) = a_d \hat{Y}_d(r) + [1 - a_d] \hat{Y}_d(syn_r)}$$

Remarques sur les 2 estimateurs BLUP d'Hidiroglou:

* Ces 2 estimateurs combinés ont un gros **avantage par rapport aux estimateurs synthétiques** (de moyenne et par le ratio) : ils ont la propriété d'être **"conservateurs".**

En effet, lorsque $n_d = N_d$ (exhaustivité dans le domaine), dans les 2 cas, on a bien $\hat{Y}_d(com_H) = Y_d$ ce qui n'est pas le cas pour les estimateurs synthétiques.

* Mais, **lorsque** $n_d << N_d$, ce qui est, en définitive le cas le plus courant, le terme $\overline{Y}_{s_d}$ estimé par $\hat{\overline{Y}}_{sd} = \overline{y}_d = \frac{\sum_{i \in sd} \frac{y_i}{p_i}}{\sum_{i \in sd} \frac{1}{p_i}}$ de chacun des estimateurs combinés est négligeable.

Le **gain des estimateurs combinés BLUP devient alors minime** par rapport aux estimateurs synthétiques associés $\hat{Y}_d(syn_m)$ et $\hat{Y}_d(syn_r)$

3 - des poids dépendant des données : 'data dependent estimators'

3 - 1 - Première démarche

Les poids optimaux pour combiner 2 estimateurs sont fonctions de l'erreur quadratique moyenne (Mean Squared Error) des estimateurs combinés et de leur covariance .

En effet, soit l'estimateur combiné suivant : $\hat{Y}_d(com) = a_d \hat{Y}_d(dir) + (1 - a_d)\hat{Y}_d(syn)$

Les poids optimaux $a_d(opt)$ sont obtenus en

$$* \min_{a_d}(MSE(\hat{Y}_d(com))$$

$$* \text{ sous la contrainte } cov(\hat{Y}_d(dir), \hat{Y}_d(syn)) = 0$$

Remarque : La contrainte est assez forte puisque grâce à elle, on tire toute l'information de l'estimateur direct et toute celle de l'estimateur synthétique sans interférence possible

On obtient :

$$a_d(opt) = \frac{MSE(\hat{Y}_d(syn))}{MSE(\hat{Y}_d(syn)) + V(\hat{Y}_d(dir))}.$$

Ces quantités sont généralement inconnues ; il faut donc **les estimer à partir des données**.

Sous l'hypothèse $cov(\hat{Y}_d(dir), \hat{Y}_d(syn)) = 0$, ces poids optimaux peuvent être estimés par

$$\hat{a}_d(opt) = \frac{(\hat{Y}_d(syn) - \hat{Y}_d(dir))^2 - V(\hat{Y}_d(dir))}{(\hat{Y}_d(syn) - \hat{Y}_d(dir))^2}$$

3 - 2 - Démarche de Purcell et Kish (79)

Ils ont cherché plutôt à minimiser la moyenne de MSE c'est à dire

$$\underset{a_d}{Min}\left\{\frac{1}{D}\sum_{d=1}^{D} MSE(\hat{Y}_d(com))\right\}$$

ce qui conduit à une estimation du poids optimal de la forme :

$$\hat{a}_d(opt) = 1 - \frac{\sum_{d=1}^{D} V(\hat{Y}_d(dir))}{\sum_{d=1}^{D} (\hat{Y}_d(syn) - \hat{Y}_d(dir))^2}$$

3 - 3 - Démarche de Fay et Herriot (79)

Réf : Robert E. Fay III et Roger A. Herriot (1979) «Estimates of income for small places : an application of James-Stein procedures to census data».

L'incertitude due à l'échantillonnage s'écrit :

$\hat{Y}_d = Y_d + e_d$ avec e_d erreur d'échantillonnage telle que

$$E(e_d) = 0$$
$$V(e_d) = v_d \text{ connue}$$
$$Cov(e_d\, e_{d'}) = 0 \;\forall d \neq d'$$

et Y_d vraie valeur

On modélise une régularité sur les Y_d supposés corrélés aux X_d sous la forme

$$Y_d = aX_d + b + u_d \text{ avec } u_d \text{ erreur de modèle telle que}$$

$$E(u_d) = 0$$
$$V(u_d) = s^2 \text{ inconnue}$$
$$Cov(u_d\, u_{d'}) = 0 \;\forall d \neq d'$$

Application :

Soit $\hat{Y}_d = \hat{Y}_d(d)$ l'estimation directe précédemment calculée dans le cas d'un sondage à Probabilité Egale Sans Remise.

Nous avons alors le modèle suivant

$$\hat{Y}_d(d) = aX_d + b + u_d + e_d$$
$$\text{avec } E(u_d + e_d) = 0$$
$$V(u_d + e_d) = V(\hat{Y}_d(d)) + s^2$$
$$Cov(u_d + e_d, u_{d'} + e_{d'}) = 0 \;\forall d \neq d'$$

On doit donc estimer a, b, et s^2 par la méthode des moindres carrés quasi généralisés. La régression fournit un estimateur qui sera noté $\hat{Y}_d(reg(s^2))$ dépendant de la variance s^2.

On sait que $E\left\{\sum_{d=1}^{D} \frac{(\hat{Y}_d(d) - \hat{Y}_d(reg(s)))^2}{V(\hat{Y}_d(d)) + s^2}\right\} = D - 2$

On utilise alors la méthode itérative de Newton pour trouver s^2. On notera alors la variance trouvée à la nième itération : $s^2(n)$

$$s^2(n+1) = s^2(n) + \frac{D - 2 - \phi(s^2(n))}{\phi'(s^2(n))}$$

$$\text{avec } \phi(s^2(n)) = \sum_{d=1}^{D} \frac{(\hat{Y}_d(d) - \hat{Y}_d(reg(s^2(n))^2}{V(\hat{Y}_d(d)) + s^2(n)}$$

L'algorithme converge rapidement en moins de 10 itérations en général vers s^2 (*).

L'estimateur de Fay-Herriot s'écrit alors comme un **estimateur combiné de l'estimateur direct et de l'estimateur de régression avec des poids fonction des variances de ces estimateurs.**

$$\hat{Y}_d(com_{FH}) = a_d \hat{Y}_d(d) + (1 - a_d)\hat{Y}_d(reg(s^2(*)))$$

$$\text{avec } a_d = \frac{s^2(*)}{s^2(*) + V(\hat{Y}_d(d))}$$

Application : la méthode de Fay-Herriot est utilisée pour répartir des subventions aux gouvernements locaux et à ceux des états des Etats-Unis à partir des résultats du Recensement de Population et des logements en 1970.

Les deux applications réalisées à partir de l'enquête Éducation 92

1 - Quelques mots sur l'enquête 'Efforts d'Education des familles'

➔ Objectifs

Cette enquête (en abrégé enquête Education) a été conçue et réalisée par l'Insee avec la collaboration de l'Ined (F. Héran et C. Gissot) en mai-juin 1992 de façon à dresser le bilan de l'année scolaire 1991-1992.

Le **thème de l'éducation** y est abordé pour la première fois à l'Insee. En effet, jusqu'ici, les enquêtes sur le système scolaire étaient réalisées auprès des établissements scolaires. Elles ne laissaient pas la parole aux parents, ni même aux enfants.

➔ Organisation de l'enquête

Le **champ de l'enquête** est constitué par les parents d'enfants âgés de 2 à 25 ans scolarisés vivant dans le ménage ou hors ménage.

Les 11703 ménages interrogés ont été sélectionnés par **un plan de sondage stratifié selon la taille de l'unité urbaine à plusieurs degrés**. Tiré dans la base de sondage de l'échantillon maître complétée par la base des logements neufs, il y a eu sur-représentation des ménages ayant déclaré un enfant de cet âge lors du recensement de 1990 et sous-représentation des logements vacants, secondaires et occasionnels.

L'enquête repose sur **plusieurs supports**

- un questionnaire 'parents' qui traite de l'attitude actuelle des parents en matière d'éducation comme le choix de l'établissement, leurs exigences envers l'établissement d'accueil, les dépenses scolaires et extra-scolaires, le temps passé auprès des enfants, les rencontres avec les enseignants.

- un questionnaire 'enfants' (collégiens, lycéens , étudiants) retraçant l'opinion de leurs enfants sur l'école.

➔ La taille de l'échantillon des répondants

Ne disposant pas d'une base de données précise sur la population des parents d'élèves âgés de 2 à 25 ans, il est normal de n'obtenir en fin de compte pour le questionnaire parents que **5265 ménages répondants**.

Mais, les responsables d'enquête ont estimé le **taux de non-réponse** ; il s'élève à 3,6% environ. Ils n'ont donc pas effectué de redressement.

➔ La population des enfants scolarisés et âgés de 2 à 25 ans

La première application porte sur la variable **niveau scolaire de l'enfant** tandis que la seconde porte sur **la dépense scolaire par enfant du ménage**. Or, la dépense scolaire par enfant du ménage tout comme le niveau scolaire de l'enfant est déclarée seulement pour au plus 2 enfants du ménage nommés A et B. Ils sont les 2 premiers parmi un classement par ordre alphabétique des enfants du ménage et hors ménage.

Par conséquent, dans les 2 cas, **un échantillon de 8292 enfants des ménages a été constitué** et les **poids** relatifs à chaque enfant des ménages en fonction des poids des ménages ont été recalculés. Ainsi, si le nombre d'enfants des ménages vivant dans le ménage et hors ménage dans le champ de l'enquête est

* égal à 1, alors le poids de l'enfant sera égal au poids du ménage

* égal à 2, les 2 enfants auront pour poids le poids du ménage

* supérieur à 2, les 2 enfants sélectionnés auront pour poids le poids du ménage multiplié par le nombre d'enfants du ménage dans le champ divisé par 2.

➔ Les publications

* Les dépenses d'éducation des familles - INSEE Première n°261 - juin 1993
* Les efforts éducatifs des familles - INSEE Résultats n°62-63 de septembre 1994
* L'aide au travail scolaire : les mères persévèrent - INSEE Première n°350 - décembre 1994

2 - Les effectifs du préélémentaire et de l'élémentaire en 91/92, par région

➔ le choix du 1er sujet d'application

La variable d'intérêt Y ici est **simple** ; c'est un effectif. D'autre part, les vrais chiffres d'effectifs issus des enquêtes effectuées auprès de tous les établissements scolaires publics et privés du 1er degré et orchestrées par l'Education Nationale étaient disponibles. Des **comparaisons** entre les résultats issus des estimations et la réalité ont pu être effectuées.

2 - 1 - L'étude

* *La population :* les enfants scolarisés pendant l'année scolaire 91/92 âgés de 2 à 25 ans

dans l'échantillon n=8292 enfants
dans la population totale N inconnu

* *Les domaines :* les régions

Nombre : D=22

N_d est le nombre d'enfants dans la population de ménages dans le domaine d

n_d dans l'échantillon

* *Les variables d'intérêt :* le niveau scolaire.

On définit 2 niveaux d'étude : le Préélémentaire et l'Elémentaire.

Soit Y_k et Z_k les variables aléatoires définies ainsi pour chaque enfant k de la population :

$$Y_k = 1 \text{ si } k \in PRE$$
$$Y_k = 0 \text{ sinon}$$

avec PRE = { enfants de la population scolarisés dans le Préélémentaire}

$$Z_k = 1 \text{ si } k \in ELE$$
$$Z_k = 0 \text{ sinon}$$

avec ELE = { enfants de la population scolarisés dans l'Elémentaire}

* *L'objectif*

On cherche à estimer des totaux des variables d'intérêt par domaine d c'est-à-dire encore le **nombre d'enfants scolarisés dans le Préélémentaire et l'Elémentaire par région** à partir des données de l'échantillon et de variables auxiliaires :

$$Y_d = \sum_{k=1}^{N_d} Y_k$$

$$Z_d = \sum_{k=1}^{N_d} Z_k$$

* *La variable auxiliaire*

On utilisera comme variable auxiliaire l'effectif réel des enfants scolarisés dans le premier degré par région. Elle est fournie par la la Direction de l'Evaluation et de la Prospective (Ministère de l'Education Nationale). On la notera X_d

* *Les vraies valeurs*

La DEP fournit aussi le détail par niveau scolaire c'est-à-dire les effectifs réels des enfants scolarisés dans le Préélémentaire et dans l'Elémentaire par région : Y_d et Z_d que l'on utilisera à des fins de validation des différentes méthodes d'estimation.

On a $X_d = Y_d + Z_d$

2 - 2 - Les estimateurs calculés

La présentation des estimateurs qui suit est faite avec le préélémentaire Y uniquement.

1 - Les estimations directes

➔ Horvitz-Thompson

$$\hat{Y}_d(h) = \sum_{i=1}^{n_d} \frac{y_i}{p_i}$$

➔ Cas particulier

Hypothèse : le tirage de l'échantillon des enfants est supposé être à probabilités égales et sans remise (PESR).

Le poids de chaque individu est donc égal à $p_i = \frac{n}{N}$

$$\hat{Y}_d(d) = \frac{N}{n} \sum_{i=1}^{n_d} y_i$$

Comme N est inconnu, **le taux de sondage** $\frac{n}{N}$ peut être estimé par le rapport du total national du 1er degré dans l'échantillon (enquête Education 92) sur le même total réel (fichier de l'Education Nationale).

$$\hat{f} = \frac{\sum_{d=1}^{D} (y_d + z_d)_{\text{échantillon}}}{\sum_{d=1}^{D} (Y_d + Z_d)_{\text{population}}}$$

➔ par le ratio estimé avec X_d effectifs réels du premier degré en 91/92 par région

$$\hat{Y}_d(rd) = X_d \tilde{R}_{Yd}$$

avec $\tilde{R}_{Yd}$ estimation de $R_{Yd} = \frac{Y_d}{X_d}$

L'estimation du ratio choisie est celle utilisant l'estimation d'Horvitz-Thompson de X_d dans 2 cas :

* *cas général :*

$$\tilde{R}_{Yd}(h) = \frac{\hat{Y}_d(h)}{\hat{X}_d(h)}$$

* *cas particulier d'un tirage PESR :*

$$\tilde{R}_{Yd}(d) = \frac{\hat{Y}_d(d)}{\hat{X}_d(d)} = \frac{\sum_{i=1}^{n_d} y_i}{\sum_{i=1}^{n_d} x_i} = \frac{y_{s_d}}{x_{s_d}}$$

2 - Les estimations synthétiques

Ce sont des **estimations synthétiques uniquement par le ratio.** L'hypothèse sous-jacente est que $R = R_d$

$$\hat{Y}_d(rsd) = X_d \tilde{R}_Y$$

avec $\tilde{R}_Y$ estimation directe de $R_Y = \frac{Y}{X}$

Là aussi on peut distinguer **2 cas** pour l'estimation du ratio :

** cas général*

$$\tilde{R}_Y(h) = \frac{\hat{Y}(h)}{\hat{X}(h)}$$

** cas particulier d'un tirage PESR*

$$\tilde{R}_Y(d) = \frac{\hat{Y}(d)}{\hat{X}(d)} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{y_s}{x_s}$$

3 - Les estimations combinées

Compte tenu du fait que N_d est inconnu, seuls les **estimateurs combinés ayant des poids soit fonction de la variable auxiliaire X, soit fonction de la variance estimée** de l'estimateur direct d'Horvitz-Thompson dans le cas d'un sondage PESR ont pu être construits. L'**estimateur de Fay-Herriot noté** $\hat{Y}_d(FH)$ aussi a pu être calculé.

1 - des poids fonction de X

Les poids sont fonction du total 1er degré X (source Education Nationale) et de son estimation d'Horvitz : $\hat{X}_d(h)$

$$a_d(h) = \frac{\hat{X}_d(h)}{X_d} = \frac{\hat{Y}_d(h) + \hat{Z}_d(h)}{Y_d + Z_d} \text{ si } a_d(h) < 1$$
$$a_d(h) = 1 \text{ si } a_d(h) \geq 1$$

D'où des estimations qui se notent : $\hat{Y}_d(cxh...)$

Dans le cas particulier d'un sondage PESR les poids s'écrivent :

$$a_d(d) = \frac{\hat{X}_d(d)}{X_d} \text{ si } a_d(d) < 1$$
$$a_d(d) = 1 \text{ si } a_d(d) \geq 1$$

D'où des estimations qui se notent : $\hat{Y}_d(cxd...)$

2 - des poids fonction de la variance estimée

Le poids optimal s'écrit - on l'a vu précédemment :

$$a_d = 1 - \frac{\sum_{d=1}^{22} V(\hat{Y}_d(d))}{\sum_{d=1}^{22} (\hat{Y}_d(syn) - \hat{Y}_d(dir))^2} \text{ si } a_d > 0$$

$$a_d = 0 \text{ si } a_d \leq 0$$

Or, on ne connaît la variance d'aucun estimateur direct que nous avons calculé. Par conséquent, il faut l'estimer.

Pour faciliter les calculs, seule la variance de l'estimateur d'Horvitz Thompson dans le cas où le tirage est à Probabilités Egales Sans Remise a été estimée (cf. 1ère partie).

Plusieurs séries d'estimateurs combinés avec des poids fonction de l'estimation de la variance ont été construits. En effet, comme les poids sont aussi fonction des estimations directes et synthétiques, on peut calculer 8 poids différents et toutes les combinaisons possibles.

Ces estimateurs seront notés $\hat{Y}_d(cvh.)$ lorsque ad est fonction de l'estimateur synthétique par le ratio d'Horvitz Thompson (cas général).
Dans le cas particulier où le tirage est PESR, ils seront notés $\hat{Y}_d(cvd.)$.

3 - l'estimateur de Fay-Herriot

$$\hat{Y}_d(FHa) = a_d \hat{Y}_d(d) + (1 - a_d)\hat{Y}_d(rega)$$

$$\text{avec } a_d = \frac{V(\hat{Y}_d(rega))}{V(\hat{Y}_d(rega)) + V(\hat{Y}_d(FHa))}$$

La régression linéaire simple associée à ce modèle s'est effectuée avec la variable indépendante

- du revenu fiscal moyen par région notée FIS

La notation utilisée est $\hat{Y}_d(FHF)$ et $\hat{Y}_d(regF)$ pour les 2 estimateurs (a=F)

- de la population scolaire totale par région notée SCO

La notation utilisée est $\hat{Y}_d(FHS)$ et $\hat{Y}_d(regS)$ pour les 2 estimateurs (a=S)

2 - 3 - Les résultats

1 - Le calage

Compte tenu de l'absence de correction des non-réponses, il s'est avéré nécessaire de **réaliser un calage** en utilisant les totaux du Préélémentaire (Y=2557940) ou de l'Elémentaire (Z=4107640) sur la France métropolitaine (fichier de l'Education Nationale).

Les formules des estimateurs deviennent alors :

$$\hat{Y}C_d(aaa) = \hat{Y}_d(aaa) \times \frac{Y}{\hat{Y}(aaa)}$$

- *Remarque sur le total du premier degré calé* :

$$\begin{aligned}\hat{X}C(aaa) &= \hat{Y}C(aaa) + \hat{Z}C(aaa) \\ &= \sum_{d=1}^{D} \left(\hat{Y}_d(aaa) \times \frac{Y}{\hat{Y}(aaa)} + \hat{Z}_d(aaa) \times \frac{Z}{\hat{Z}(aaa)}\right) \\ &= Y + Z = X\end{aligned}$$

Les estimations du 1er degré obtenues aprés calage correcpondent au total réel X, par construction.

2 - Les statistiques de comparaison

Soit l'erreur relative de l'estimateur calé du Préélémentaire

$$EC(\hat{Y}C_d(aaa)) = \frac{\hat{Y}C_d(aaa) - Y_d}{Y_d}$$

avec Y_d vraie valeur

Pour comparer les différentes estimations entre elles, 2 statistiques ont été retenues : **la moyenne et l'écart-type de la valeur absolue de l'écart relatif entre l'estimateur et la vraie valeur, calculés sur les 22 régions.**

3 - Les résultats

1 - Au niveau national n=8292 enfants

Niveau scolaire	Effectif réel Y Z X	Effectif échantillon y z x	Effectif estimé $\hat{Y}(h)$ $\hat{Z}(h)$ $\hat{X}(h)$
Préélémentaire Y	2557940	1546	2318024
Elémentaire Z	4107640	2395	3575087
Premier degré X	6665580	3941	5893111

Grâce à ce tableau, on constate l'utilité d'un calage.

2 - Le Préélémentaire

Les estimateurs combinés avec des poids fonction de la variance sont ici les meilleurs. Sur l'ensemble des 22 régions, ils ont les moyennes les plus faibles et leurs écarts-types sont aussi très faibles.

- **Le classement des estimateurs selon les statistiques de comparaison**

Le **tableau n°1** (p. 341) fournit le classement de tous les estimateurs calculés selon la moyenne sur les 22 régions. On trouvera en tête les **estimateurs qui combinent l'estimateur d'Horvitz Thompson** dans le cas particulier d'un sondage PESR qui seul n'est pourtant pas très performant **avec un estimateur synthétique**. Ce dernier apporte sûrement beaucoup de sa performance.

On peut aussi remarquer que ce sont ces estimateurs combinés qui ont des poids fonctions des estimateurs par le ratio (direct et synthétique).

Viennent ensuite en 5ème position du point de vue de la moyenne, les 2 **estimations synthétiques par le ratio.**

Les **estimateurs combinés dont les poids sont fonction de X** ne sont pas performants puisque le meilleur du point de vue de la moyenne atteint déjà 5,967 et 5,089 en écart-type.

L'estimateur de Fay-Herriot n'est pas non plus 'bon'.

* En utilisant comme variable auxiliaire le revenu fiscal moyen par région, la moyenne de l'écart en valeur absolue s'élève à 11,023 et son écart-type à 9,061 à cause d'une corrélation entre les 2 variables moyenne (0,753) et de la très mauvaise performance de l'estimateur de régression associé. L'estimateur de

Fay-Herriot tire par conséquent sa maigre performance uniquement de l'estimateur direct.

* La variable auxiliaire constituée de la population scolaire totale par région apporte un peu plus d'information que la précédente à l'estimateur de Fay-Herriot (Corr=0,983).

- **L'annexe 1** fournit

☐ les valeurs de quelques estimateurs des effectifs du préélémentaire en 91/92 sur les 22 régions comme
- l'estimateur d'Horvitz Thompson
- l'estimateur synthétique par le ratio
- le meilleur (du point de vue de la moyenne de l'indicateur de comparaison) estimateur combiné dont le poids est fonction de la variance

☐ les écarts associés

☐ ainsi que la taille de l'échantillon par région n_d.

Cette dernière information permet d'avoir un regard différent sur les résultats précédents puis qu'elle permet de constater le rôle important joué par n_d dans les estimations région par région.
En particulier, pour la Corse avec 32 enfants dans l'échantillon, on constate une forte diminution de l'écart relatif à l'estimateur direct à l'écart relatif à l'estimateur synthétique ou combiné (de 41 à 6).

Dans ce tableau, on peut constater à nouveau la bonne performance domaine par domaine de l'estimateur combiné dont le poids est fonction de la variance estimée.

Tableau n°1 : Classement des différents estimateurs d'effectifs d'élèves scolarisés dans le Préélémentaire en 91/92 selon la moyenne de l'écart relatif en valeur absolue

Type	Estimations régionales CALEES	Moyenne	Ecart-type
CV	$a_d(rd,rsd)\hat{Y}_d(d) + (1 - a_d(rd,rsd))\hat{Y}_d(rsd)$	**2,146**	1,767 (4)
CV	$a_d(rd,rsd)\hat{Y}_d(d) + (1 - a_d(rd,rsd))\hat{Y}_d(rsh)$	**2,152**	1,763 (3)
CV	$a_d(rd,rsh)\hat{Y}_d(d) + (1 - a_d(rd,rsh))\hat{Y}_d(rsd)$	**2,156**	1,760(2)
CV	$a_d(rd,rsh)\hat{Y}_d(d) + (1 - a_d(rd,rsh))\hat{Y}_d(rsh)$	**2,162**	1,757 (1)
S	$\hat{Y}_d(rsh)$ et $\hat{Y}_d(rsd)$	**2,286**	1,833
CV	...(56)		
FH	regSCO	**5,471**	7,813
CX	$a_d(Xh)\hat{Y}_d(rd) + (1 - a_d(Xh))\hat{Y}_d(rsh)$	**5,967**	5,089
CX	$a_d(Xh)\hat{Y}_d(rd) + (1 - a_d(Xh))\hat{Y}_d(rsd)$	**5,969**	5,092
CX	...(2)		
D	$\hat{Y}_d(rd)$	**6,962**	5,608
FH	FHSCO	**7,222**	7,556
CX	...(4)		
D	$\hat{Y}_d(rh)$	**8,747**	6,823
CX	...(2)		
CV	...(3)		
CX	...(2)		
CV	...(1)		
FH	FHFIS	**11,023**	9,061
D	$\hat{Y}_d(d)$	**11,474**	8,883
CX	...(4)		
D	$\hat{Y}_d(h)$	**16,661**	9,556
FH	regFIS	**72,487**	71,462

La légende adoptée dans les 2 tableaux est la suivante :

* pour le **type** d'estimateur :

- D comme direct
- S comme synthétique
- CX comme combiné avec un poids de combinaison fonction de la variable auxiliaire X
- CV comme combiné avec un poids de combinaison fonction de la variance estimée
- FH comme Fay-Herriot

* pour signaler l'existence d'estimateurs (sans en indiquer leur moyenne et écart-type) et de leur nombre m, on notera leur type commun et...**(m)**

3 - L'Elémentaire

- **Les estimations réalisées pour l'Elémentaire sont globalement meilleures que celles réalisées pour le Préélémentaire.**

- **Le classement des estimateurs selon les statistiques de comparaison est globalement le même que celui trouvé pour le Préélémentaire. Les estimateurs combinés avec des poids fonction de la variance ainsi que les estimateurs synthétiques viennent toujours en tête.**

- **Le classement des estimateurs selon les statistiques de comparaison (tableau n°2 p. 344)**

Il faut noter la très bonne performance de **l'estimateur de Fay Herriot** avec comme variable auxiliaire le total premier degré DEG due sans doute à l'excellente corrélation entre la variable d'intérêt et la variable auxiliaire (0,994).

Ainsi même si on dispose de peu d'informations : un estimateur direct tout à fait médiocre et une variable auxiliaire très fortement corrélée mais disponible seulement en agrégat sur la région, on obtient un estimateur de Fay Herriot tout à fait honorable ainsi qu'un estimateur de régression.

On trouvera en tête **les estimateurs synthétiques** ainsi que tous **les estimateurs combinés ayant des poids de combinaison fonction de la variance** et de Yrd et de Yrsd. Ces poids a_d sont nuls ce qui entraîne l'égalité de ces estimateurs avec les estimateurs synthétiques.

Tout comme pour le Préélémentaire, on peut remarquer que ce sont les estimateurs combinés dont les poids sont fonction des estimateurs par le ratio (direct et synthétique) qui leur succèdent mais ils combinent l'estimateur direct par le ratio - et non pas le cas particulier d'Horvitz Thompson - avec un estimateur synthétique.

L'estimateur de Fay-Herriot avec pour variable auxiliaire l'effectif scolaire du premier degré suit.

Les estimateurs combinés dont les poids sont fonction de X sont là encore peu performants puisque le meilleur du point de vue de la moyenne atteint déjà 3,739 et 3,171 en écart-type.

- **L'annexe 2** fournit

☐ les valeurs de quelques estimateurs des effectifs de l'élémentaire en 91/92 sur les 22 régions comme
- l'estimateur d'Horvitz Thompson
- l'estimateur synthétique par le ratio
- le meilleur (du point de vue de la moyenne de l'indicateur de comparaison) estimateur combiné dont le poids est fonction de la variance et différent du précédent

☐ les écarts associés.

☐ ainsi que la taille de l'échantillon par région

Tableau n°2 : Classement des différents estimateurs d'effectifs d'élèves scolarisés dans l'élémentaire en 91/92 selon la moyenne de l'écart relatif en valeur absolue

Type	Estimations régionales CALEES	Moyenne	Ecart-type
S	$\hat{Z}_d(rsh)$ et $\hat{Z}_d(rsd)$	**1,380**	1,062
CV	...(16)	**1,380**	1,062
CV	$a_d(rh,rsh)\hat{Z}_d(rd) + (1 - a_d(rh,rsh))\hat{Z}_d(rsh)$	**1,460**	1,084
CV	$a_d(rh,rsh)\hat{Z}_d(rd) + (1 - a_d(rh,rsh))\hat{Z}_d(rsd)$	**1,461**	1,086
CV	$a_d(rh,rsd)\hat{Z}_d(rd) + (1 - a_d(rh,rsd))\hat{Z}_d(rsh)$	**1,472**	1,096
CV	$a_d(rh,rsd)\hat{Z}_d(rd) + (1 - a_d(rh,rsd))\hat{Z}_d(rsd)$	**1,474**	1,098
CV	(24)		
FH	regDEG	**2,303**	2,130
FH	FHDEG	**2,561**	2,665
CV	(6)		
CX	$a_d(Xh)\hat{Z}_d(rd) + (1 - a_d(Xh))\hat{Z}_d(rsd)$	**3,739**	3,171
CX	$a_d(Xh)\hat{Z}_d(rd) + (1 - a_d(Xh))\hat{Z}_d(rsh)$	**3,740**	3,170
CV	...(6)		
CX	...(2)		
D	$\hat{Z}_d(rd)$	**4,345**	3,477
CV	...(4)		
CX	...(2)		
FH	regSCO	**4,952**	7,703
CX	...(2)		
D	$\hat{Z}_d(rh)$	**5,524**	4,256
FH	FHSCO	**8,845**	6,306
CX	...(2)		
CV	...(3)		
CX	...(2)		
CV	...(1)		
CX	...(4)		
FH	FHFIS	**11,107**	7,761
D	$\hat{Z}_d(d)$	**11,424**	7,469
D	$\hat{Z}_d(h)$	**12,70**	9,640
FH	regFIS	**63,427**	56,249

3 - La dépense scolaire par enfant scolarisé dans le premier degré en 91/92 par région

3 - 1 - L'étude

* *La population :* les enfants scolarisés pendant l'année scolaire 91/92 dans **le premier degré**

dans l'échantillon n=3941 enfants
dans la population totale N=6665580 enfants

* *Les domaines :* les régions

Nombre : D=22

N_d est le nombre d'enfants scolarisés dans le premier degré dans la population de ménages dans le domaine d

n_d dans l'échantillon

* *La variable d'intérêt :* la dépense scolaire.

Elle est définie comme la somme des frais de pension - 1/2 pension (PENSI), d'inscription et d'assurances (INASS), des achats de fournitures et de vêtements scolaires (FOVET), des dépenses de loisirs et de sorties dans le cadre de l'école (LOSOR), d'achats de livres scolaires (LIVSC), et de titres de transports (TRANS).

DESCOL=PENSI+INASS+FOVET+LOSOR+LIVSC+TRANS

* *L'objectif*

On cherche à estimer **la dépense scolaire par enfant du premier degré et selon la région d** c'est à dire encore

$$\frac{1}{N_d}\sum_{k=1}^{N_d} DESCOL_k$$

avec $DESCOL_k$ la dépense scolaire totale de l'enfant du premier degré k

et N_d l'effectif d'enfants du premier degré dans la population.

** Les méthodes*

Deux solutions s'offrent alors à nous compte tenu de la formule :

1 - **estimer le total** de la dépense scolaire totale par domaine $\sum_{k=1}^{N_d} DESCOL_k$ puis diviser cette estimation par **Nd**

Car, contrairement à la première application, N_d cette fois ci est connu ; c'est l'effectif d'enfants scolarisés dans le premier degré dans la population totale disponible dans le fichier du ministère de l'Education Nationale.
On notera ces estimations $\hat{MD}_d$

2 - ou bien **estimer le tout** en prenant en compte par conséquent une estimation de Nd. Ces estimations seront notées $\hat{DM}_d$

** L'absence de vraies valeurs*

Contrairement à l'application précédente, nous ne disposons pas des vraies valeurs des dépenses scolaires par enfant scolarisé dans le premier degré.

3 - 2 - Les estimateurs calculés

1 - Les estimations directes

➔ Horvitz-Thompson

1 - La dépense scolaire par enfant connaissant Nd

$$\hat{MD}_d(h) = \frac{\hat{D}_d(h)}{N_d}$$

$$\text{avec } \hat{D}_d(h) = \sum_{i=1}^{n_d} \frac{DESCOL_i}{P_i}$$

2 - On peut aussi l'estimer plus directement par la moyenne

$$\hat{DM}_d(h) = \frac{\hat{D}_d(h)}{\sum_{i=1}^{n_d} \frac{1}{P_i}}$$

➔ **Cas particulier : le plan de sondage est PESR**

1 - Le poids affecté à chaque enfant est constant et est égal à 1/f. Ainsi la dépense scolaire par enfant peut être estimée par

$$\hat{MD}_d(D) = \frac{\hat{D}_d(d)}{N_d}$$

$$\hat{D}_d(d) = \frac{1}{f}\sum_{i=1}^{n_d} DESCOL_i$$

$$\text{avec } f = \frac{n}{N}$$

2 - Mais dans l'hypothèse PESR, on sait que le meilleur estimateur sans biais d'une moyenne est la moyenne sur l'échantillon.

$$\hat{DM}_d(d) = \frac{\sum_{i=1}^{n_d} DESCOL_i}{n_d} = \frac{\hat{D}_d(d)}{n_d}$$

➔ **Post-stratifié**

1 -

$$\hat{MD}_d(pst) = \frac{\hat{D}_d(pst)}{N_d} = \hat{DM}_d(h)$$

$$\text{car } \hat{D}_d(pst) = N_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL_i}{p_i}}{\sum_{i=1}^{n_d} \frac{1}{p_i}} = N_d \frac{\hat{D}_d(h)}{\hat{N}_d(h)}$$

2 - On peut aussi directement estimer la moyenne

$$\hat{DM}_d(pst) = N_d \frac{\hat{DM}_d(h)}{\hat{N}_d(h)}$$

➔ **Post stratifié groupé**

1 - Les groupes sont les 2 niveaux scolaires du premier degré

$$\hat{MD}_d(pst_g) = \frac{\hat{D}_d(pst_g)}{N_d}$$

$$\text{avec } \hat{D}_d(pst_g) = Y_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL * Y_i}{p_i}}{\sum_{i=1}^{n_d} \frac{Y_i}{p}} + Z_d \frac{\sum_{i=1}^{n_d} \frac{DESCOL * Z_i}{p_i}}{\sum_{i=1}^{n_d} \frac{Z_i}{p_i}}$$

en adoptant les notations précédentes Y pour le préélémentaire et Z pour l'élémentaire

2 - On peut là-aussi directement estimer la moyenne à partir des 2 estimations de moyenne sur les 2 groupes

$$\hat{DM}_d(pst_g) = Y_d \frac{\hat{DYM}_d(h)}{\hat{Y}_d(h)} + Z_d \frac{\hat{DZM}_d(h)}{\hat{Z}_d(h)}$$

en notant $\hat{DYM}_d(h)$ et $\hat{DZM}_d(h)$ les estimations d'Horvitz-Thompson des moyennes des dépenses scolaires respectivement pour le Préélémentaire et pour l'Elémentaire.

2 - Les estimations synthétiques

Elles sont sans biais si l'hypothèse de la moyenne est vérifiée

$$\overline{DM_d} = \overline{DM}$$

➔ **simple**

1 -

$$\hat{MD}_d(sn) = \frac{\hat{D}_d(sn)}{N_d} = \hat{DM}(h)$$

$$\text{avec } \hat{D}_d(sn) = N_d \frac{\hat{D}(h)}{\hat{N}(h)}$$

Ainsi la moyenne synthétique revient à celle calculée par Horvitz-Thompson au niveau national et son écart-type sur les 22 régions sera nul.

2 - autre méthode:

$$\hat{DM}_d(sn) = N_d \frac{\hat{DM}(h)}{\hat{N}(h)}$$

➔ **groupé**

1 - Les groupes sont les niveaux scolaires du premier degré

$$\hat{MD}_d(sn_g) = \frac{1}{N_d}(Y_d \frac{\hat{DY}(h)}{\hat{Y}(h)} + Z_d \frac{\hat{DZ}(h)}{\hat{Z}(h)})$$

2 - autre méthode:

$$\hat{DM}_d(sn_g) = Y_d \frac{\hat{DYM}(h)}{\hat{Y}(h)} + Z_d \frac{\hat{DZM}(h)}{\hat{Z}(h)}$$

3 - Les estimations combinées

1 - des poids fonction de la taille du domaine

Soit le poids $a_d = \frac{\hat{N}_d(h)}{N_d}$ si $a_d \leq 1$

$a_d = 1$ sinon

1 - Première méthode :

On peut alors construire 8 estimateurs combinés compte tenu des 4 estimateurs directs du total : $\hat{D}_d(h)\ \hat{D}_d(d)\ \hat{D}_d(pst)\ \hat{D}_d(pstg)$ et des 2 estimateurs synthétiques $\hat{D}_d(sn)$ et $\hat{D}_d(sn_g)$. Les estimateurs sont alors notés : $\hat{D}_d(cNa)$

Puis on calcule les moyennes connaissant Nd

$$\hat{MD}_d(cNa) = \frac{\hat{D}_d(cNa)}{N_d}$$

avec $a \in [1,8]$

2 - Deuxième méthode

Les estimateurs de la combinaison ne sont pas les dépenses scolaires totales mais déjà des moyennes.

Ainsi, on notera

$$\hat{D}M_d(cNa) = a_d\hat{D}M_d(dir) + (1-a_d)\hat{D}M_d(syn)$$
$$\text{avec } a \in [1,8]$$

2 - les poids fonction de la variance estimée

1 - Première méthode :

L'idée ici est la même que dans l'application précédente. On part du poids optimal

$$a_d = 1 - \frac{\sum_{d=1}^{22} V(\hat{D}_d(dir))}{\sum_{d=1}^{22} (\hat{D}_d(syn) - \hat{D}_d(dir))^2} \text{ si } a_d > 0$$
$$a_d = 0 \text{ si } a_d \leq 0$$

On estime la variance dans le cas le plus simple où l'estimateur direct est l'estimateur d'Horvitz Thompson dans le cas où le tirage est PESR.

Plusieurs séries d'estimateurs combinés avec des poids fonction de l'estimation de la variance peuvent alors être construits.

Ces estimateurs seront notés $\hat{D}_d(cvh.)$ lorsque ad est fonction de l'estimateur synthétique par le ratio d'Horvitz Thompson (cas général).

Enfin, on divise ces dépenses scolaires totales par Nd

$$M\hat{D}_d(cv..) = \frac{\hat{D}_d(cvh.)}{N_d}$$

2 - Deuxième méthode :

La variance qui figure dans le poids est celle d'une estimation directe de la dépense scolaire moyenne. Je l'ai estimée ainsi :

$$V(\hat{D}M_d(d)) = \frac{V(\hat{D}_d(d))}{N_d^2}$$

Le poids optimal est alors estimé par

$$a_d = 1 - \frac{\sum_{d=1}^{22} \hat{V}(\hat{D}M_d(d))}{\sum_{d=1}^{22} (\hat{D}M_d(syn) - \hat{D}M_d(dir))^2} \text{ si } a_d > 0$$

$$a_d = 0 \text{ si } a_d \leq 0$$

Les estimateurs combinés de ce type seront alors notés :

$$\hat{D}M_d(cv..)$$

3 - l'estimateur d'Hidiroglou

Soit le poids $a_d = \frac{n_d}{N_d}$

1 - 1ère méthode :

$$\hat{D}_d(BLUP) = a_d \hat{D}_d(pst) + (1 - a_d)\hat{D}_d(sn)$$

$$M\hat{D}_d(BLUP) = \frac{\hat{D}_d(BLUP)}{N_d}$$

Les poids de la combinaison sont trés faibles. Ainsi,

$$M\hat{D}_d(BLUP) \approx M\hat{D}_d(sn)$$

2 - 2ème méthode :

$$\hat{D}M_d(BLUP) = a_d \hat{D}M_d(pst) + (1 - a_d)\hat{D}M_d(sn)$$

4 - l'estimateur de Fay-Herriot

1 - 1ère méthode :

$$\hat{D}_d(FHa) = a_d \hat{D}_d(d) + (1 - a_d)\hat{D}_d(rega)$$

$$\text{avec } a_d = \frac{V(\hat{D}_d(rega)}{V(\hat{D}_d(rega)) + V(\hat{D}_d(FHa))}$$

- avec le revenu fiscal moyen par région
La notation utilisée est $\hat{D}_d(FHF)$ et $\hat{D}_d(regF)$ pour les 2 estimateurs (a=F)

- avec les effectifs d'enfants scolarisés dans le public par région
La notation utilisée est $\hat{D}_d(FHP)$ et $\hat{D}_d(regP)$ pour les 2 estimateurs (a=P)

Ensuite on calcule la moyenne en divisant par Nd : $\hat{DM}_d(FHP)$

2 - 2ème méthode :

avec la proportion d'effectifs d'enfants scolarisés dans le public par région

$$\hat{DM}_d(FHP) = a_d\hat{DM}_d(d) + (1-a_d)\hat{DM}_d(regP)$$

$$\text{avec } a_d = \frac{V(\hat{DM}_d(regP))}{V(\hat{DM}_d(regP)) + V(\hat{DM}_d(d))}$$

3-3 - Les résultats

1 - Au niveau national

Les dépenses scolaires par enfant estimées sont les suivantes :

Niveau scolaire	$M\hat{D}(d) = \hat{DM}(d)$	$M\hat{D}(h)$	$\hat{DM}(h)$
Préélémentaire	974,6	936,7	848,8
Elémentaire	1473,3	1427,3	1242,3
Premier degré	1277,7	1234,3	1091,3

Les intervalles de confiance des estimations des dépenses scolaires par enfant estimées dans le cas d'un tirage pesr sont les suivants :

Niveau scolaire	<	$M\hat{D}(d) = \hat{DM}(d)$	<
Préélémentaire	884,6	974,6	1064,6
Elémentaire	1398,5	1473,3	1548,2
Premier degré	1219,6	1277,7	1335,8

2 - Au niveau régional

Comment comparer les estimateurs entre eux ne connaissant pas la dépense scolaire des enfants du premier degré réelle?

- comparer leur moyenne ou leur écart-type sur les 22 régions ?

- regarder les différences entre toutes les estimations région par région ?

Le tableau n°3 fournit la moyenne sur les 22 régions des différents estimateurs

Tableau n°3 : Classement des différents estimateurs de la dépense scolaire par enfant du premier degré en 91/92 selon leur moyenne empirique sur les 22 régions

Type	Estimations régionales	Moyenne
D	$\hat{MD}_d(h)$	964,44
D	$\hat{MD}_d(d) = \frac{\hat{D}_d(d)}{N_d}$	1111,06
D	$\hat{DM}_d(h) = \hat{MD}_d(pst)$	1121,32
D	$\hat{MD}_d(pst_g)$	1133,02
D	$\hat{DM}_d(d)$	1141,60
D	$\hat{DM}_d(pst)$	1337,23
D	$\hat{DM}_d(pst_g)$	2599,71
S	$\hat{MD}_d(sn) = \hat{DM}(h)$	1234,33
S	$\hat{MD}_d(sn_g)$	1240,96
CN	$\frac{a_d\hat{D}_d(dir) + (1-a_d)\hat{D}_d(syn)}{N_d}$	de 1017,10 à 1147,09
C	$\hat{MD}_d(BLUP)$	1234,27
CV	$\frac{a_d\hat{D}_d(dir) + (1-a_d)\hat{D}_d(syn)}{N_d}$	de 1171,70 à 1224,66
FH	$\hat{MD}_d(FHP) = \frac{a_d\hat{D}_d(d) + (1-a_d)\hat{D}_d(regF)}{N_d}$	1096,06
FH	$\hat{MD}_d(FHF) = \frac{a_d\hat{D}_d(d) + (1-a_d)\hat{D}_d(regF)}{N_d}$	1093,17
FH	$\hat{DM}_d(FHP) = a_d\hat{DM}_d(d) + (1-a_d)\hat{DM}_d(regP)$	1166,96

Conclusion

Il est bien évident que **le sujet n'est pas épuisé** et que de nombreux estimateurs sont encore à appliquer ne serait ce que sur cette enquête comme les estimateurs de Battese, Harter et Fuller, les estimateurs temporels et l'estimateur hiérarchique de Bayes.

D'autres études sur de nouvelles enquêtes-ménages mais aussi sur des enquêtes-entreprises peuvent être envisagées...

Je vois dans **le logiciel POULPE de calcul de précision** développé par l'Unité Méthodes Statistiques une ouverture intéressante puisqu'il permettrait à la fois de calculer les variances des estimateurs nationaux dans différents cas de sondage plus élaborés que le tirage PESR et d'utiliser ces formules de variance plus appropriées dans le cas des estimateurs combinés. De plus, il pourrait estimer les variances des estimateurs sur des domaines ce qui permettrait une comparaison

En effet, la partie qui me semble intéressante à développer à ce stade de la recherche est celle concernant **la comparaison des performances** de ces différents estimateurs. Cette comparaison de précision peut être faite sur l'ensemble des domaines ou domaine par domaine...

Singh, Gambino et Mantel souligne dans leur dernière étude l'importance de mesurer les performances des estimateurs sur chacun des domaines.

> *'L'élaboration de méthodes qui permettraient d'estimer l'Erreur Quadratique Moyenne pour des domaines pris individuellement devrait figurer parmi les priorités de recherche.'*

L'autre idée intéressante soulignée par **Singh, Gambino, Mantel** est de **prévoir toutes les utilisations d'une enquête avant d'élaborer son plan d'échantillonnage** et par là-même les estimations sur petits domaines. Ainsi, les plans de sondage des grandes enquêtes doivent être établis de telle sorte que les données inférées sur des domaines préétablis qualifiés de 'planifiés' soient fiables. Un arbitrage entre la nécessité de recourir à l'estimation pour domaine et le désir d'obtenir une certaine efficience aux niveaux d'agrégation supérieurs s'impose alors.

> *'On devrait prendre conscience de la question des petites régions dés le début de la conception des plans de sondage pour les grandes enquêtes'*
>
> Singh, Gambino, Mantel
>
> 'Les Petites Régions : Problèmes et Solutions' (94)

BIBLIOGRAPHIE

'Updating Small Area Population Estimates in England and Wales', Stephen Simpson, Ian Diamond, Pete Tonkin - *Royal Statistical Society* (1996).

'Robust Estimation of Mean Squared Error of Small Area Estimators', P. Lahiri, J.N.K. Rao - *Journal of the American Statistical Association*, vol.90 n°430 (juin 1995).

'Generalized sample size dependent estimators for small areas', A.C. Singh, *I.U.H.* Mian - ARC'95.

'Small Area Estimation at Provincial Level in the Italian Labour Force Survey', Pieo D. Falarsi, Stefano Falorsi, Aldo Russo - *ARC 95*.

'Comparaison empirique de méthodes d'estimation pour petites régions pour l'enquête sur la population active italienne', P.D. Falorsi, S. Falorsi, A. Russo - *Techniques d'enquête*, vol20 n°2 pp179-184 (déc 1994).

'Borrowing Strength from past data in Small Domain Prediction by Kalman Filtering - A Case Study', Arijit Chaudhuri, Tapabrata Maiti - 1994.

'Small Area estimation : an Appraisal', M.Ghosh, J.N.K. Rao - *Statistical Science*, vol.9, n°1 - 2 (1994).

'Estimating Activity limitation in the noninstitutionalized population : a method for small areas' - J. Elston Lafata,G.G. Koch, W.G. Weissert - *American Journal of Public Health*, vol.84, n°11 pp1813-1817 (nov.1994).

'Les petites régions : problèmes et solutions', M.P. Singh, J. Gambino et H.J. Mantel - *Techniques d'enquête*, vol.20, n°1 pp3-23 (juin 1994).

'MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête', A.C. Singh, H.J. Mantel, B.W. Thomas - *Techniques d'enquête*, vol.20, n°1 pp35-46 (juin 1994).

'Estimation pour petits domaines dans des plans de sondage avec probabilités inégales', D. Holt, D.J. Holmes - *Techniques d'enquête*, vol.20, n°1 pp25-33 (juin 1994).

'Quelques aspects particuliers des sondages - Estimation sur des domaines', P. Ardilly - *Les sondages*, chapitre IV, Techniques de sondage (1994).

'An application of small area estimation techniques to derive state estimates of health insurance coverage from the 1987 nmes', J.J. Braden, B. Cohen - *Journal of Economic and Social Measurement*, 20, pp193-213 (1994).

'Small Area Estimation : Overview and Empirical Study', J.N.K. Rao, G.H. Choudhry - *ICES Proceedings* (1993).

'Estimation for domains', C.E. Särndal, B. Swensson, J. Wretman - *Model Assisted Survey Sampling*, chapitre 10 pp386-417 Ed.Springer-Verlag (1992).

'Issues and options in the provision of small area data', Singh, Gambino, Mantel - *International Conference on Small Area Statistics and Survey Designs* - Warsaw (sept-oct 1992).

'Design-based approaches in estimation for domains', C.E. Särndal - *International Conference on Small Area Statistics and Survey Designs* - Warsaw (sept-oct 1992).

'Estimation pour les petits domaines : théorie et pratique à Statistique Canada', M.A. Hidiroglou - *Actes des Journées de Méthodologie Statistique 13 et 14 mars 1991*- INSEE-Méthodes n°29-30-31 pp375-401 (déc.1992).

'Méthode d'utilisation d'enquête à un niveau géographique où l'échantillon est faible', F. Jeger - *Actes des Journées de Méthodologie Statistique 13 et 14 mars 1991* - INSEE-Méthodes n°29-30-31 pp363-373 (déc.1992).

'Bayesian prediction in linear models : Applications to small area estimation', G.S.Datta, M. Ghosh - *The Annals of Statistics*, vol.19, n°4 pp1748-1770 (1991).

'Estimation de la production de blé par comté', E.A. Stasny, P.K. Goel, D.J. Rumsey - *Techniques d'enquête*, vol.17, n°2 pp229-244 (décembre 1991).

'Evaluation of procedures for improving population ; Estimates for small areas', K.M.Wolter, B.D. Causey - *Journal of the American Statistical Association*, vol.86, n°414 (juin 1991).

'Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales', D. Pfeffermann, L. Burck - *Techniques d'enquête*, vol.16, n°2 pp229-249 (décembre 1990).

'The Estimation of the Mean Squared Error of Small-Area Estimators', N.G.N. Prasad, J.N.K. Rao - *Journal of the American Statistical Association*, vol.85, n°409 (March 1990).

'A Bayesian Approach to Small Domain Estimation', Kung-Jong Lui and William G. Cumberland - *Journal of Official Statistics*, vol.5, n°2 (1989).

'Small Domain Estimation : A Conditional Analysis', Carl-Erik Särndal and Michael A. Hidiroglou - *Journal of the American Statistical Association*, vol.84, n°405 (March 1989).

Evaluation of small area estimators : an empirical study, G.H. Choudhry and J.N.K. Rao - (1988), Central Bureau of Statistics OSLO.

'Application of some empirical bayes methods to small area statistics', Emil Spjotvoll and Ib Thomsen - *Invited Paper 29.2 46 th Session of the ISI.*

'An error components model for prediction of county crop areas using survey and satellite data', George E. Battese, Rachel M. Harter, Wayne A. Fuller - april 14 (1986).

'Synthetic estimators, SPREE and best model-based predictors of small area means', J.N.K. Rao - 1986.

'Synthetic estimates for small areas : problems and results of a simulation experiment', Adam Marton - *Journal of the United Nations ECE 4*, 71-80 North-Holland (1986).

'Small area statistics an international symposium', R. Platek, J.N.K. Rao, C.E. Särndal, M.P. Singh - Ottawa May 1985.

'Regression analysis and ratio analysis for domains : a randomization-theory approach', Eva Elvers, Carl Erik Särndal, Jan H. Wretman, Göran Örnberg - *Journal of Statistics*, vol.13, n°2 (1985) (La Revue Canadienne de Statistique).

'An overview of small area estimation techniques', J. Dumais, S. Earwaker, JF. Gosselin, D. Paton, KP. Srinath, R. Verma - march (1985).

'The multivariate components of variance model for small area estimation', Wayne A. Fuller, Rachel M. Harter - November 4 (1985).

'Design-consistent versus model-dependent estimation for small domains', Carl Erik Särndal - *Journal of the American Statistical Association*, vol.79, n° 387, september (1984).

'Une bibliographie pour l'estimation pour les petites régions', Jean Dumais, David Paton, Ravi Verma, Stephen Earwaker, J.F. Gosselin, K.P. Srinath - *Techniques d'enquête*, vol.9, n°2 (1983).

'Postcensal estimates for local areas using current samples with census as the source of sampling frame', S.M. Tam - *International Statistical Review* (1982).

'On estimating population and income for local areas', Evelyn M. Kitagawa and Bruce D. Spencer - May 1981.

'Postcensal estimates for local areas (or domains)', Noel J. Purcell and Leslie Kish - *International Statistical Review* (1980).

'A biometrics invited paper', Noel J. Purcell and Leslie Kish - *Biometrics* 35, 365-384, june 1979.

'Estimates of income for small places : an application of james-stein procedures to census data', Robert E. Fay III and Roger A. Herriot - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'A model-Based approach to estimation for small subgroups of a population', D. Holt, TMF. Smith, and T.J. Tomberlin - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'A predictive approach to subdomain estimation in finite populations', Petter Laake - *Journal of the American Statistical Association*, vol.74, n°366, june 1979.

'Some estimators for domain totals', M.P. Singh and R. Tessier - *Journal of the American Statistical Association*, vol.71, n°354, june 1976.

'A regression method for estimating population changes of local areas', Eugene P. Ericksen - *Journal of the American Statistical Association*, vol.69, n°348, december 1974.

'Estimation for domains in multistage sampling', Myint tin and Than toe - *Journal of the American Statistical Association*, vol.67, n°340, december 1972.

RG	n_d échantillon	Y_d vraie valeur	$\hat{Y}_d(h)$ direct	$\hat{Y}_d(rsh)$ synthétique	$\hat{Y}_d(cvdl7)$ combiné	$Ec(\hat{Y}_d(h))$ écart	$Ec(\hat{Y}_d(rsh))$ écart	$Ec(\hat{Y}_d(cvdl7))$ écart
11	1615	487364	446043.35	483263.53	484031.10	8.4784	0.84136	0.68386
21	203	63936	71080.40	63897.27	63998.76	11.1743	0.06058	0.09816
22	289	89583	111296.82	91073.16	91217.81	24.2388	1.66344	1.82491
23	247	80502	89666.63	85834.54	85970.87	11.3844	6.62411	6.79346
24	305	101864	77021.98	103944.64	104109.73	24.3874	2.04256	2.20464
25	198	64105	79468.77	65099.57	65202.97	23.9666	1.55147	1.71276
26	196	66430	81412.39	67980.41	68088.38	22.5536	2.33390	2.49643
31	667	220591	270700.44	213483.25	213822.33	22.7160	3.22214	3.06843
41	334	106471	93979.11	106971.30	107141.21	11.7327	0.46990	0.62947
42	259	68324	59819.16	71820.24	67871.53	12.4478	5.11715	0.66224
43	185	50275	47962.55	51464.42	51546.16	4.5996	2.36583	2.52842
52	517	146545	188179.84	146856.34	147089.60	28.4110	0.21246	0.37162
53	395	133330	120505.70	127630.68	127833.39	9.6185	4.27460	4.12256
54	215	65941	56242.95	66273.85	66379.12	14.7072	0.50478	0.66441
72	316	107771	90903.63	110456.17	110631.61	15.6511	2.49155	2.65434
73	337	96271	111114.46	93982.01	94131.28	15.4184	2.37765	2.22260
74	89	24206	16642.17	24746.35	24785.65	31.2478	2.23229	2.39466
82	781	256818	235810.10	251678.24	252077.98	8.1801	2.00132	1.84567
83	196	49956	42690.43	50998.16	51079.16	14.5439	2.08616	2.24830
91	321	89745	93497.48	89513.20	89655.38	4.1813	0.25828	0.09986
93	595	178514	168378.44	180995.83	181283.30	5.6777	1.39027	1.55131
94	32	9398	5523.19	9976.82	9992.67	41.2301	6.15902	6.32763

moyenne	16,66	2,29	2,15
écart-type	9,56	1,83	1,77

ANNEXE 2

Quelques estimateurs des effectifs de l'élémentaire en 91-92

RG	n_d échantillon	Z_d vraie valeur	$\hat{Z}_d(h)$ direct	$\hat{Z}_d(rsh)$ synthétique	$\hat{Z}_d(cvh31)$ combiné	$EC(\hat{Z}_d(h))$ écart	$EC(\hat{Z}_d(rsh))$ écart	$EC(\hat{Z}_d(cvh31))$ écart
11	1615	771943	739084.76	776043.47	776492.16	4.2566	0.53119	0.58931
21	203	102570	92762.79	102608.73	102668.06	9.5615	0.03776	0.09560
22	289	147739	155452.02	146248.84	146333.40	5.2207	1.00864	0.95141
23	247	143169	141005.79	137836.46	137916.15	1.5109	3.72465	3.66898
24	305	168999	134761.50	166918.36	167014.87	20.2590	1.23115	1.17405
25	198	105534	88944.17	104539.43	103487.12	15.7199	0.94242	1.93955
26	196	110716	93333.50	109165.59	107966.49	15.7001	1.40035	2.48339
31	667	335712	410002.67	342819.75	343017.96	22.1293	2.11722	2.17626
41	334	172279	142687.52	171778.70	171878.02	17.1765	0.29040	0.23275
42	259	118828	135987.22	115331.76	115398.44	14.4404	2.94227	2.88616
43	185	83833	80468.02	82643.58	82691.36	4.0139	1.41880	1.36180
52	517	236139	262596.43	235827.66	235964.01	11.2042	0.13185	0.07411
53	395	199255	204981.51	204954.32	205072.83	2.8740	2.86032	2.91979
54	215	106758	118495.84	106425.15	106486.68	10.9948	0.31178	0.25415
72	316	180060	165000.02	177374.83	177477.39	8.3639	1.49126	1.43431
73	337	148631	158821.42	150919.99	151007.25	6.8562	1.54005	1.59876
74	89	40279	32879.03	39738.65	39761.63	18.3718	1.34151	1.28447
82	781	399015	358178.35	404154.76	404388.43	10.2344	1.28811	1.34667
83	196	82937	85571.51	81894.84	81942.19	3.1765	1.25657	1.19948
91	321	143512	181171.88	143743.80	143826.91	26.2416	0.16152	0.21943
93	595	293132	316040.42	290650.17	290818.22	7.8151	0.84666	0.78933
94	32	16600	9413.63	16021.18	16030.44	43.2914	3.48690	3.43109

moyenne	12,70	1,380	1,46
écart-type	9,64	1,062	1,08

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 :

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

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

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 :

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

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 :

$$ECI_S(n,z) = | \; TC_S(n,z) - TCI(n,z) | \; .$$

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 $ECI_S(n,z) < a1 \; NO_S$, où $a1$ est un paramètre à choisir (voisin de 2), on ne modifie pas W_S, poids a priori de S. Autrement dit, si $WMI_S(n,z)$ est le coefficient de modulation de W_S (coefficient compris entre 0 et 1), on prend $WMI_S(n,z) = 1$;

- si $ECI_S(n,z) > b1 \; NO_S$, 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 : $WMI_S(n,z) = 0$;

- si $a1 \; NO_S \leq ECI_S(n,z) \leq b1 \; NO_S$, on interpole $WMI_S(n,z)$ en fonction de la valeur de $ECI_S(n,z)$:

$$WMI_S(n,z) = (b1 \; NO_S - ECI_S(n,z)) / ((b1-a1) \; NO_S).$$

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 : $WI_S(n,z) = W_S \; WMI_S(n,z)$.

7.2.2. Itérations.

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

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

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. Δ(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*.

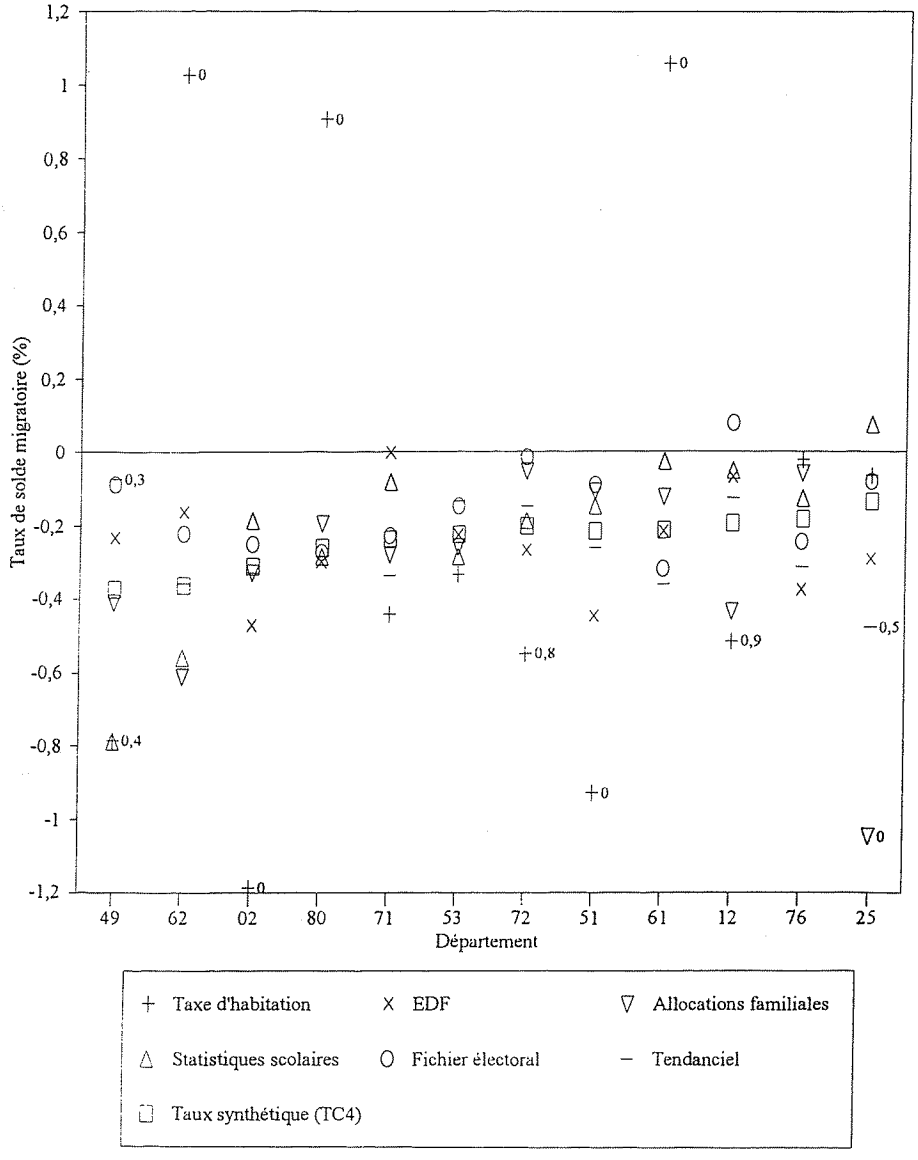


Figure 1 : Synthèse des taux de solde migratoire de l'année 1990 pour douze départements, repérés par leur numéro (49, 62...).

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.

LE ZONAGE EN AIRES URBAINES : UNE NOUVELLE APPROCHE DE LA VILLE ET DE SON ESPACE PÉRIURBAIN

Thomas Le Jeannic

Cet article a pour but de préciser la méthode de travail et les divers développements d'idées qui ont prévalu pour définir le zonage en aires urbaines (ZAU), zonage devant remplacer les zones de peuplement industriel ou urbain (ZPIU).

Il n'a pas pour but d'être exhaustif, et les nombreuses idées intéressantes qui ont été émises, durant deux années de fonctionnement du groupe de travail, ne figurent pas toutes dans ce document.

Il est organisé en sept parties :

1 - Structuration passée du territoire

2 - Investigations auprès d'experts internes et externes à l'Insee

3 - Principes de base pour un nouveau zonage

4 - Les pôles urbains

5 - Aires urbaines, communes périurbaines

6 - La multipolarité

7 - L'espace à dominante rurale.

1. Structuration passée du territoire

La France, comme le reste du monde, voit les populations se concentrer autour des grands pôles urbains. Ce phénomène, souvent désigné sous le terme de métropolisation, peut être défini comme l'« exercice de forces centripètes conduisant à la concentration des activités et des hommes dans les espaces urbains les plus peuplés, tandis que les villes moyennes et les espaces ruraux perdent, au moins relativement, de la vitalité »[1]

Dans le même temps, les centres urbains ne pouvant accueillir indéfiniment de la population et l'automobile permettant une dédensification ou une plus grande dispersion du tissu urbain, ce phénomène s'accompagne de celui de péri-urbanisation, voire de rurbanisation, c'est-à-dire d'un débordement du centre vers la périphérie. Ces forces centrifuges concernent aussi bien les grandes villes que les villes moyennes et même petites.

Pour analyser ces phénomènes complexes et partiellement contradictoires, nous disposions jusqu'à présent sur l'ensemble du territoire métropolitain de deux nomenclatures de zonage :

- les unités urbaines,
- les zones de peuplement industriel ou urbain (ZPIU)

Les unités urbaines : une première approche de la ville

La ville était plus facile à définir au Moyen-Âge,. Elle rassemblait un nombre plus ou moins important d'habitations au milieu d'un océan de verdure et de champs. Elle était cernée d'une muraille.

Plus tard, la révolution industrielle a engendré un exode rural massif. Avec l'intense phénomène d'urbanisation qui l'accompagna, la ville a grossi et débordé au dehors de ses murailles. Les territoires communaux ne se sont pas pour autant agrandis, sauf exception comme Paris ou Metz qui ont absorbé leurs faubourgs. Est alors apparue la nécessité de définir des agglomérations, fondées sur la continuité de l'habitat et pouvant contenir plusieurs communes. On a ainsi défini des **unités urbaines**, au sein desquelles on a distingué la ville-centre et les communes de

[1] G.F. DUMONT

banlieue. La ville-centre correspond en général à la ville historique, parfois même encore entourée de ses murailles.

La délimitation des unités urbaines *(annexe 3)* :

- On identifie tout d'abord sur le territoire les zones bâties, susceptibles d'atteindre 2 000 habitants au recensement de 1990.

- Une **zone bâtie** est constituée par des constructions avoisinantes formant un ensemble tel qu'aucune ne soit séparée de la plus proche de plus de 200 mètres. Les terrains utilisés à des fins publiques tels que jardins publics, aérodromes, routes ... ; ceux utilisés à des fins industrielles ou commerciales tels qu'usines, magasins, voies ferrées ..., ainsi que les cours d'eau traversés par des ponts ne sont pas pris en compte lors de la détermination de la distance séparant les habitations.

- On s'est tout d'abord intéressé aux zones bâties qui s'étendaient sur deux ou plusieurs communes. Parmi les communes concernées, on a éliminé celles dont la population appartenant à la zone bâtie représentait moins de la moitié de la population de la commune. Si les communes non éliminées étaient au moins au nombre de deux et si la partie de leur population qui résidait dans la zone atteignait au total au moins 2 000 habitants, ces communes constituaient une **agglomération multicommunale**.

- À l'issue du recensement, les communes qui n'appartiennent pas à une agglomération multicommunale délimitée préalablement au recensement, ont été classées comme urbaines au sens de l'Insee lorsque le nombre d'habitants de la plus grande zone bâtie (au sens précédemment défini) de la commune atteignait au moins 2 000 habitants. Ces communes urbaines sont également appelées **villes isolées**.

- Finalement, on a appelé unités urbaines aussi bien les agglomérations multicommunales que les villes isolées. Au sens de l'Insee, toute commune appartenant à une unité urbaine est réputée « urbaine », toutes les autres communes étant classées « rurales ».

Reposant sur la notion de continuité de l'habitat, cette définition sous-entendait à l'origine une vision très tranchée et relativement peu nuancée de l'espace : d'un côté, les communes urbaines représentant « la ville », de l'autre, les communes rurales, « la campagne ».

Mais très vite, cette vision s'est avérée caduque. Avec les développements parallèles de l'automobile et de la maison individuelle, de nouveaux modes de vie sont apparus. Des citadins sont venus habiter des communes apparemment rurales, tout

en conservant de fréquents contacts avec la ville. Cet entremêlement de l'habitat rural et du mode de vie urbain a rendu plus floues les limites de la ville. Pour en tenir compte, l'Insee a proposé dès les années soixante, une nouvelle définition en complément de celle des unités urbaines : les **zones de peuplement industriel ou urbain (ZPIU).**

Les zones de peuplement industriel ou urbain (ZPIU) : une vision plus extensive de l'urbanisation

Les ZPIU sont composées de trois types de communes :

a) des unités urbaines

Toutes les unités urbaines appartiennent à une ZPIU, et chacune doit appartenir à une seule ZPIU. Toutefois, une ZPIU peut ne comprendre aucune unité urbaine, ou en comprendre une ou plusieurs.

b) des communes industrielles

Une commune rurale a été classée comme industrielle, si elle comptait un ou plusieurs établissements industriels et commerciaux ou administratifs (chantiers de bâtiment et des travaux publics mis à part) de 20 salariés au moins, à la condition toutefois que l'effectif total de ces établissements dépassât 100 salariés.

c) des communes dortoirs

Les communes-dortoirs sont tout d'abord les communes rurales non industrielles répondant à la condition suivante:

(% sortants) > 1,2x(% ménages agricoles)

où % sortants = poucentage d'actifs résidents travaillant hors de la commune

% ménages agricoles = poucentage de ménages ordinaires vivant de l'agriculture.

Exemples :

pour 20 % de ménages agricoles, il suffisait de 24 % de sortants ;
pour 50 % de ménages agricoles, 60 % de sortants.

- De cette façon, on sélectionne les communes qui ont un faible taux d'agriculteurs ou/et une forte proportion de migrants alternants. C'est bien ce que l'on entendait par « communes-dortoirs ». Le problème est que les informations nécessaires à ce calcul ne sont disponibles que plusieurs mois, voire plusieurs années après le déroulement du recensement. Afin de ne pas retarder d'autant la nouvelle délimitation des ZPIU et d'avoir une définition du périmètre disponible dès la parution des premiers résultats, on évalue cette formule à l'aide de variables observées lors du recensement précédent (1954 pour les premières ZPIU dites « 1962 »,..., 1982 pour les ZPIU dites « 1990 »).

- Pour pallier en partie cet inconvénient, une formule de « rattrapage » permet de récupérer les communes « presque » dortoirs, dès lors qu'elles ont un taux de croissance démographique élevé. Elles sont probablement devenues communes-dortoirs depuis le dernier recensement. On décide donc de retenir les communes répondant à la condition suivante :

(% sortants) > 1,2x(% ménages agricoles) - 1,1(% variation de population)

où % variation population = pourcentage d'évolution de la population entre les deux derniers recensements.

Exemple :
pour 20 % de ménages agricoles et 10 % d'augmentation de population, il faut seulement 13 % de sortants.

Au cours des années soixante et soixante-dix, les ZPIU ont permis de mesurer et de décrire les principales caractéristiques des phénomènes de métropolisation et de périurbanisation. Cependant, victimes de leur « succès », elles ont petit à petit été atteintes de gigantisme, ainsi que le montrent la carte 1 de même que le tableau suivants :

Évolution des zones de peuplement industriel ou urbain (ZPIU)

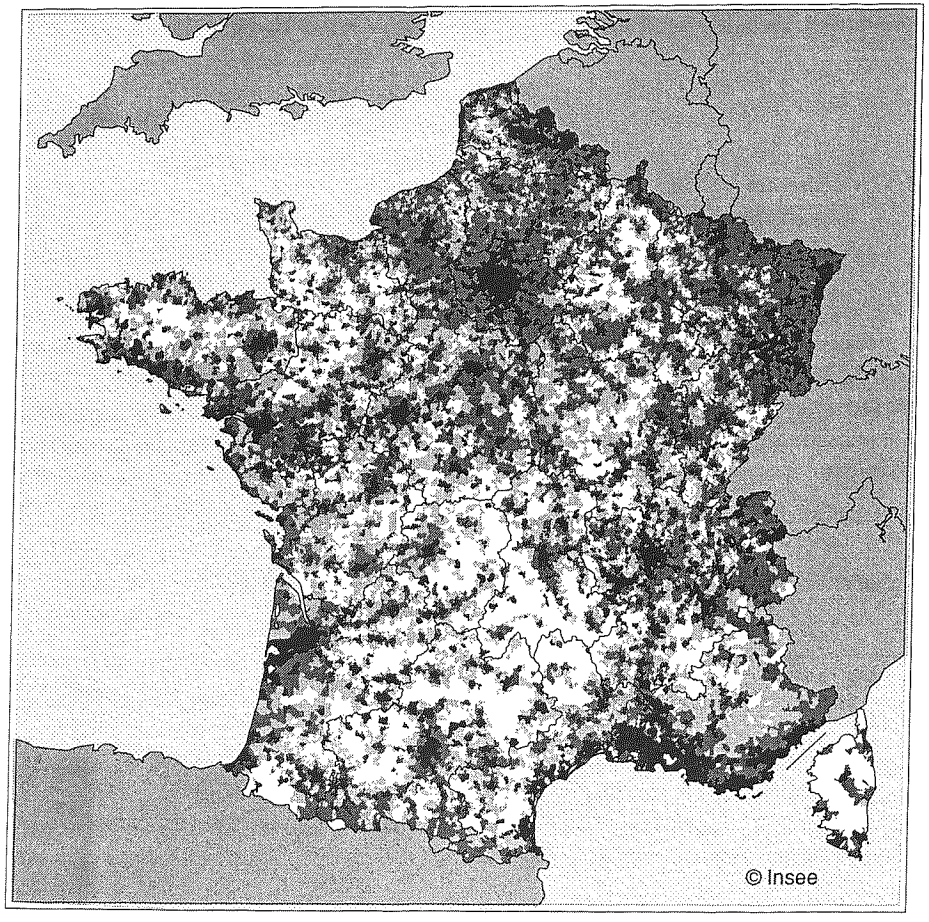
Recensement	Nombre de communes en ZPIU	% des communes en ZPIU	% de la population en ZPIU
1962	9 100	24	74
1975	12 000	33	83
1982	19 000	52	89
1990	28 500	78	96

Source : Insee - Recensement de la population

Rappelons que les « ZPIU 1990 » sont calculées sur la base des résultats du recensement de 1982 : la prise en compte des résultats du recensement de 1990 accentuerait encore ce phénomène, laissant probablement moins de deux millions d'habitants hors ZPIU. Dès lors, l'appartenance à une ZPIU n'est plus un critère discriminant : très utile, pendant trente ans, cette notion semble avoir atteint aujourd'hui sa limite.

Face à ce constat, il devenait nécessaire de définir un nouveau concept spatial capable de fournir une grille de lecture pertinente du territoire, cette grille devant être utilisée pour diffuser les résultats du prochain recensement de la population.

Carte 1-
Les zones de peuplement industriel ou urbain (ZPIU)
Évolution de 1982 à 1990



2. Investigation auprès d'experts internes et externes à l'Insee

Fin 1992, un questionnaire fut envoyé à toutes les directions régionales de l'Insee pour connaître leurs besoins en matière de zonage de type urbain/rural, ainsi que leurs critiques de fond sur la décomposition alors en vigueur.

Deux besoins se sont alors exprimés. Le premier, par l'ensemble des directions régionales, portait sur la prise en compte de l'aire d'influence des centres urbains, sur leur environnement immédiat. Un moyen de mesure généralement proposé de cette influence était les flux domicile-travail, en y incluant éventuellement des informations sur les équipements provenant de l'inventaire communal.

Le deuxième besoin, exprimé surtout par des régions à fort caractère rural, était une meilleure connaissance du monde rural, qu'il soit dynamique ou en voie de désertification.

Les critiques qui furent émises sur les ZPIU concernaient :

- l'utilisation contestable, dans les formules, d'une variable de dynamisme démographique (% de la variation de population), variable qui pouvait ensuite être utilisée pour décrire le zonage et produire des tautologies du style : les ZPIU, par essence dynamiques, sont dynamiques ;

- la problématique des pôles qui n'était pas exprimée clairement. La délimitation des unités urbaines d'une part, et du reste des ZPIU d'autre part, était en effet traitée de manière relativement distincte, sans lien apparent. Les migrations alternantes étaient utilisées uniquement en niveau, sans tenir compte de leurs directions ;

- le fait d'utiliser des données de recensements précédents pour le découpage de 1990, qui n'est plus tolérable aujourd'hui ;

- sans être d'une très grande complexité, la définition des ZPIU n'était pas des plus simples à présenter. Il faut reconnaître que peu de personnes, y compris à l'Insee, connaissaient par cœur leur définition précise. Le nouveau zonage devrait non seulement être plus pertinent, mais défini de façon suffisamment simple pour que les nombreux utilisateurs potentiels se l'approprient plus aisément.

Courant 1993, une série d'entretiens fut menée auprès de 26 personnes, dont la moitié travaille à l'Insee (*voir en annexe 1, la liste des interviewés*). Les personnes de l'extérieur appartenaient à des organismes tels que le Centre National de

de l'extérieur appartenaient à des organismes tels que le Centre National de Recherche Scientifique (CNRS), l'Institut National de Recherche Agronomique (INRA), la SEGESA, des universités, la Délégation à l'Aménagement du Territoire et à l'Action Régionale (DATAR), la Fédération Nationale des Agences d'Urbanisme (FNAU), le SCEES.

Il s'agissait alors de recueillir le maximum d'idées, de pistes de recherche pour un nouveau zonage.

Un certain nombre de points ont été abordés :

Problèmes méthodologiques généraux :

- Deux types de démarches sont possibles pour décrire le territoire :

On peut réaliser soit une typologie de l'espace, soit un zonage relationnel. Dans une typologie, les éléments d'une même classe sont associés parce qu'ils se ressemblent en fonction de certains critères. Le résultat est alors généralement un morcellement du territoire.

Dans un zonage relationnel, les éléments d'une zone sont associés parce qu'ils ont des liens entre eux. Le lien est défini par l'intensité des échanges.

Certains pensent que ces deux démarches sont incompatibles, d'autres qu'il est possible de les raccorder.

- Il apparaît absolument nécessaire, comme pour les ZPIU, de pouvoir suivre le zonage au cours du temps. Cela implique une contrainte sur le choix des outils et des critères à utiliser.

- Dans la description du territoire, plusieurs options sont possibles dans le choix du concept ou de l'indicateur. On peut raisonner sur la population résidente, sur les emplois et les activités économiques, sur les équipements, sur l'occupation du sol ... Est-il possible de trouver un zonage unique, capable de répondre à ces différentes optiques ?

Distinction entre les notions d'urbanisation et d'agglomération

Comme on l'a vu précédemment, la mesure de l'urbanisation s'est faite jusqu'à présent à partir de la notion d'unité urbaine, c'est-à-dire d'agglomération (au moins

2 000 habitants, pas plus de 200 mètres ...). Or la plupart des interlocuteurs s'accordent pour dissocier les deux concepts d'urbanisation et d'agglomération. C'est qu'ils estiment qu'il existe des unités urbaines, relativement peu peuplées, qu'il faudrait classer dans le rural, car elles présentent des activités économiques dont la caractéristique est d'être tournée vers la société rurale environnante. Ces agglomérations rurales se caractérisent par une situation de dépendance hiérarchique vis-à-vis d'agglomérations plus importantes et par la présence d'un nombre restreint d'activités économiques (commerce de détail, services administratifs et enseignement). On y trouve tout de même nombre de sous-préfectures dont c'est la fonction essentielle, ce qui limite leur dépendance hiérarchique.

Ainsi, nombre d'auteurs, dès lors qu'ils veulent analyser les villes françaises, commencent par imposer un seuil minimal de population. L'inconvénient, c'est que ce seuil varie d'un auteur à l'autre. Exprimé en général en nombre, ce seuil est pour certains à 5 000 habitants, souvent à 10 000 habitants mais aussi à 20 000 habitants pour d'autres. Enfin, quelques uns considèrent qu'un seuil unique n'est pas probant, car ce seuil dépend de l'environnement. Une unité urbaine de 10 000 habitants ne joue sans doute pas le même rôle suivant qu'elle se trouve dans une région très rurale, ou à proximité d'une grande métropole.

Comment appréhender le périurbain ?

D'une manière générale, le territoire périurbain correspond à des espaces peu denses qui satisfont à une demande d'espace par les acteurs urbains : ménages pour se loger ou pour des activités de loisirs, et entreprises pour s'étendre.

Mais où arrêter les limites du périurbain, et notamment, faut-il restreindre le champ du périurbain par rapport aux ZPIU ? Quelques experts considèrent que ce qui prédomine, c'est l'extension continuelle et irréversible de la ville. L'explosion des ZPIU et leur recouvrement quasi total du territoire n'est dès lors plus aberrant. La société actuelle, avec ses moyens modernes de communication et de transport, a tendance en effet à uniformiser les comportements sur l'ensemble du territoire. Vouloir à tout prix distinguer un rural sans influence de la ville, serait un combat d'arrière garde.

Cependant, l'avis général va dans le sens d'un espace périurbain plus restreint que la notion de commune-dortoir des ZPIU. Il convient en effet de distinguer dans l'espace rural, un espace à dominante agricole qui, sans pouvoir être qualifié de « rural profond », ne saurait se définir comme essentiellement périurbain.

Pour mesurer le périurbain, les migrations alternantes (ou domicile-travail) sont considérées comme une variable fondamentale. Il est souvent recommandé de prendre en compte uniquement les migrants alternants vers le pôle, comme c'est le

cas dans les autres pays européens. Cependant, certaines communes périurbaines ne sont pas polarisées car elles sont attirées par plusieurs villes.

Le taux de migrants alternants peut parfois être trompeur, dans le cas de communes où la part des inactifs est importante. Une condition supplémentaire en valeur absolue est peut-être nécessaire. Enfin, le nombre de migrants n'est pas indépendant de la superficie de la commune, or la superficie moyenne des communes est variable d'une région à l'autre.

D'autres variables ont été proposées : données de l'inventaire communal, flux téléphoniques, importance de l'habitat individuel, poids de la construction neuve, existence d'un marché foncier de terrains à bâtir et dynamique de ce marché.

Ces différentes variables doivent permettre d'appréhender la demande d'espace des ménages. Il semble par contre plus difficile de caractériser la demande d'espace des entreprises.

Partitionner le rural hors périurbain

L'utilisation des ZPIU amenait à décomposer le monde rural en deux catégories : le rural en ZPIU représentait le périurbain ; le rural hors ZPIU était qualifié de « rural traditionnel » par les plus respectueux, de « rural profond » par les autres.

Cette partition du rural en deux catégories est généralement considérée comme insuffisante. Il semble nécessaire de faire des distinctions dans le rural non périurbain, qui n'est pas homogène. De plus, dans le concept de ZPIU, le rural est défini de manière résiduelle par rapport à l'urbain, et le rural profond de manière résiduelle par rapport au périurbain. Ceci est souvent considéré comme non satisfaisant et certains insistent sur la nécessité d'une définition positive du rural.

Plusieurs propositions de découpage du rural hors périurbain ont été faites :

- Le rural dynamique se caractérise par l'importance de l'agriculture en tant qu'utilisatrice de l'espace. La part de la superficie agricole utilisée, pourrait en être une mesure (mais attention aux conséquences de la politique agricole commune (PAC)) ;

- le rural attractif correspond aux besoins du tourisme et aux implantations de résidences secondaires et de retraités. Les espaces touristiques seraient cependant difficiles à classer, certains étant ruraux, d'autres urbains.

- les bassins industriels et agro-alimentaires, reprennent les critères des communes industrielles des ZPIU, en tenant compte des superficies, pour les

bassins industriels, des IAA et des gros agriculteurs pour les bassins agro-alimentaires ;

- les espaces délaissés se caractérisent par un excès de disponibilité d'espace par rapport à la demande des acteurs économiques. L'utilisation des densités de peuplement est souvent recommandée comme critère de partition. Certains recommandent de prendre les densités lissées, pour tenir compte de l'hétérogénéité du maillage communal ; d'autres font remarquer que la densité n'a de signification qu'en terme relatif par rapport à la densité de l'espace environnant, et qu'elle fait peut-être la part trop belle à l'existence ou à l'absence de forêts.

Délimitation des zones

Dans la démarche des ZPIU on commençait par déterminer les communes rurales susceptibles d'appartenir à une ZPIU. Ensuite seulement, on distinguait les zones les unes par rapport aux autres ; les pôles n'apparaissaient donc pas déterminants dans la définition du zonage.

Cette démarche pourrait être inversée. Il faudrait d'abord définir les pôles, puis agréger les communes ayant des liens suffisamment importants avec le pôle.

Comme on l'a vu, ces pôles ne seraient pas forcément toutes les unités urbaines. Ils pourraient être déterminés par le niveau de population, la densité ou encore la concentration des emplois, le taux d'emploi.

De plus, il pourrait y avoir différents niveaux de pôles, dans un système hiérarchisé. A côté des pôles urbains pourraient être définis des pôles ruraux. Plus que les migrations alternantes dans ce cas-là, le rayonnement serait mesuré par les données de l'inventaire communal. Il s'agirait alors d'un rayonnement de services. Cependant, la notion de polarisation ne serait pas très importante dans le rural, car les relations entre le bourg-centre et l'espace rural environnant ne sont pas nécessairement très intenses. On distinguerait alors dans l'espace rural communes équipées et communes non équipées.

La description de l'espace urbain et périurbain relèverait ainsi plutôt d'un zonage polarisé, tandis que celle de l'espace rural tendrait plutôt vers une typologie. La mise en cohérence des deux n'est pas évidente.

Ces entretiens ont débouché sur des conclusions variées et parfois contradictoires. Ils ont cependant permis d'éclaircir grandement le débat et de poser quelques jalons. Suite à ces entretiens, il a été convenu de retenir un certain nombre d'orientations.

La construction du nouveau zonage en remplacement des ZPIU pouvait se présenter de la façon suivante :

a/ des pôles à déterminer en premier = unités urbaines répondant à des critères de population ou d'emploi ;

b/ du périurbain sous l'influence de ces pôles = communes polarisées par ces pôles ; le critère de polarisation étant certainement les migrations alternantes et/ou les flux constatés par l'inventaire communal ;

c/ du rural dynamique au sein de l'espace rural = communes déterminées à partir de critères à tester ;

d/ le reste = communes n'appartenant pas aux trois catégories précédentes.

3. Principes de base pour un nouveau zonage

Un groupe de travail s'est constitué début 1994, sous l'égide des divisions « recensement » et « statistiques et études régionales ». Il était composé uniquement de personnes de l'Insee, représentant des directions régionales ou de la direction générale (*voir en annexe 2 la liste des participants*). Les réunions et nombreuses réflexions se sont déroulées sur deux années.

La méthode de travail a été très empirique. Diverses méthodes ont été testées avec des variables et des seuils différents. A chaque fois, les résultats obtenus étaient confrontés à la connaissance du « terrain » qu'avaient les participants. Il s'agissait avant tout de définir des villes et leur espace périurbain. Le seuil de définition de l'urbain est déjà quelque peu subjectif. Le seuil de définition de l'espace périurbain l'est encore davantage. On a en réalité plus affaire à un continuum qu'à une pure dichotomie.

Les participants avaient donc conscience que, quelle que soit la méthode employée, on aboutirait à une définition arbitraire et peut-être inadaptée dans certains cas précis. C'est le propre de toute nomenclature. Définir une nomenclature spatiale était bien l'objectif du groupe de travail. Au cours des premières réunions, quelques principes ont été retenus auxquels devait répondre le zonage final :

1. C'est un zonage d'étude. La nomenclature obtenue ne respecte donc aucune limite administrative, si ce n'est celles des communes et du territoire national.

2. Une unité urbaine ne peut être dissociée. Toutes les communes qui la forment sont affectées en bloc à un espace ou à une zone.

Les premiers tests ont été réalisés sur l'ensemble des communes Mais il s'est avéré que des communes d'une même unité urbaine pouvaient être classées dans des catégories différentes. Il nous a paru plus judicieux de conserver l'unité de chaque agglomération. La continuité de l'habitat est une donnée physique qu'il serait dommage de ne pas respecter.

3. La notion de commune industrielle a été abandonnée.

La notion de commune industrielle ne semble plus très pertinente en tant que telle. La France a connu un déclin industriel important depuis une quinzaine d'années, signe d'une profonde mutation de l'économie. De plus, la notion de secteur industriel elle-même recouvre des types de fonction très différentes selon que l'on se trouve en espace rural ou urbain.

4. La référence à la population active agricole a été également abandonnée.

La population active agricole a été divisée par six durant les cinquante dernières années. Elle est devenue très minoritaire, y compris dans le monde rural. Elle différencie de plus en plus mal les différents types de communes au sein de celui-ci.

5. L'aspect urbain a été traité de façon prioritaire.

Le phénomène primordial aujourd'hui est le phénomène urbain et tout ce qui gravite autour. Les trois quarts de la population française sont urbains (au sens des unités urbaines) et le dernier quart est de plus en plus dépendant de l'économie urbaine. L'essentiel des moteurs de l'économie et en tout cas de l'emploi se situe dans les villes. On a tenté de décrire le territoire en observant la façon dont il était occupé par la population. Il paraissait donc cohérent de définir d'abord les espaces les plus peuplés, les plus dynamiques démographiquement, les villes et leurs banlieues, pour analyser ensuite le reste du territoire.

6. Constitution des aires urbaines autour de pôles déterminés à priori.

Deux démarches différentes ont été envisagées par le groupe :

1). On laisse s'agglomérer des communes entre elles sans définir a priori des pôles. Et c'est l'ampleur de la zone obtenue, le niveau de sa population, qui indiquent l'existence éventuelle d'un pôle urbain. Ce type de démarche est possible avec des logiciels de zonage tels que « MIRABELLE » ou « ZONAGE ».

2). La définitio n des pôles urbains précède celle de l'espace périurbain.

S'il n'y a pas eu unanimité sur ce sujet, le choix s'est porté sur la seconde solution consistant à définir ce qu'est une « ville » pour ensuite mesurer son aire d'attraction.

7. Le critère unique de mesure de l'attraction urbaine est constitué par les migrations alternantes.

L'utilisation des migrations alternantes a été évoquée par la plupart des spécialistes du sujet.

8. Mise en évidence d'espaces connexes plus vastes contenant plusieurs aires urbaines. Dans de tels ensembles, les communes périurbaines peuvent être attirées par plusieurs de ces aires.

Dans une conurbation, où plusieurs agglomérations sont relativement proches, des communes rurales ou urbaines peuvent envoyer globalement dans celles-ci une forte proportion d'actifs, sans qu'elles soient très attirées par l'un ou l'autre de ces pôles. Ces communes sont tout autant périurbaines que des communes fortement attirées par un seul pôle.

Le groupe de travail a remis ses premières conclusions et propositions dans un rapport à la fin 1994. Ce rapport a été soumis à l'ensemble des directions régionales. Un certain nombre de critiques ont été émises, plus particulièrement sur la décomposition de l'espace rural (voir chapitre 7. L'espace à dominante rurale). Le groupe s'est dont remis au travail pour proposer une nouvelle définition présentée dans un rapport final au début de l'été 1995. Il a modifié ses définitions concernant l'espace à dominante urbaine, mais a surtout restreint ses ambitions en ne proposant plus de décomposition de l'espace à dominante rurale. Le nouveau zonage en aires urbaines était né.

Il peut se présenter de deux manières :

en typologie

1. Espace à dominante urbaine
    1.1. Pôles urbains
    1.2. Communes périurbaines
        a) Couronnes périurbaines
        b) Communes multipolarisées
2. Espace à dominante rurale

en zones emboîtées

1. Espace à dominante urbaine
    1.1 Espace urbain
        1.1.1 Aire(s) urbaine(s)
            a) Pôle urbain
            b) Couronne périurbaine
        1.1.2. Communes multipolarisées
2. Espace à dominante rurale

Nous allons maintenant passer en revue les différentes catégories d'espace définies, en précisant à chaque fois quelques unes des hypothèses testées, et la définition finalement adoptée.

4. Les pôles urbains

A travers les pôles urbains c'est bien la notion de ville qu'il s'agissait d'approcher. Le terme de ville n'a cependant pas été retenu car trop usité dans le langage courant. Il aurait certainement conduit à des utilisations impropres.

Pour définir la ville, le mieux était pour le groupe de faire des hypothèses et de comparer le résultat à la connaissance concrète des villes qu'avaient les participants.

Une ville peut se définir de diverses manières :

- **démographiquement**. Elle est un lieu de rassemblement important de population sur une superficie restreinte.

- **morphologiquement**. C'est un ensemble conséquent de constructions serrées dédiées au logement des hommes et à leurs activités.

- **économiquement**. Dans le temps, elle était le lieu de développement du commerce, de l'artisanat, de toutes sortes d'activités qui s'épanouissaient dès lors qu'un surplus agricole des campagnes le permettait. Aujourd'hui, elle serait le lieu d'activités du tertiaire supérieur.

- **administrativement**. Elle est le lieu d'exercice du pouvoir. Elle a donc ses représentations administratives.

- **en terme d'équipement**. Elle possède un lycée, un hôpital, une caisse de sécurité sociale, un hypermarché.

- culturellement.

Le but du groupe était de trouver une définition simple pouvant grosso modo résumer les différents aspects d'une ville. Lourde tâche.

Le plus simple, au départ, a été de passer en revue les différents seuils de population et de voir les problèmes qu'ils posaient.

Ces seuils de population ont été appliqués aux unités urbaines. Rappelons que cette notion d'unité urbaine n'a pas été abandonnée. Elle représente une réalité morphologique qu'il est intéressant de suivre au cours du temps. Mais elle n'apparaît plus comme condition suffisante pour définir le concept de ville.

Tout le monde s'est accordé pour dire qu'en dessous de 5 000 habitants, une unité urbaine ne peut raisonnablement pas être considérée comme une ville. Au-dessus de 20 000 habitants, on ne prend guère de risque à les qualifier toutes ainsi. La difficulté a été de trouver un seuil intermédiaire.

Un seuil trop élevé présente l'inconvénient d'éliminer des unités urbaines qui possèdent les principales fonction urbaines qui localement jouent un rôle important. Les cas de Figeac et Foix en Midi-Pyrénées, ou Dourdan en Ile-de-France, ont été cités parmi d'autres. De plus, certaines zones d'emploi ont des pôles qui sont des unités urbaines de moins de 10 000 habitants. Il serait gênant d'avoir trop de zones d'emploi sans pôle urbain.

Un seuil plus faible présente quant à lui l'inconvénient de classer comme ville des grosses unités urbaines, certes peuplées, mais qui ne remplissent essentiellement que la fonction résidentielle. C'est le cas de grosses banlieues urbaines situées à la périphérie d'agglomérations très importantes comme Paris ou Lyon. Une part importante de leur population peut quotidiennement travailler dans la mégapole.

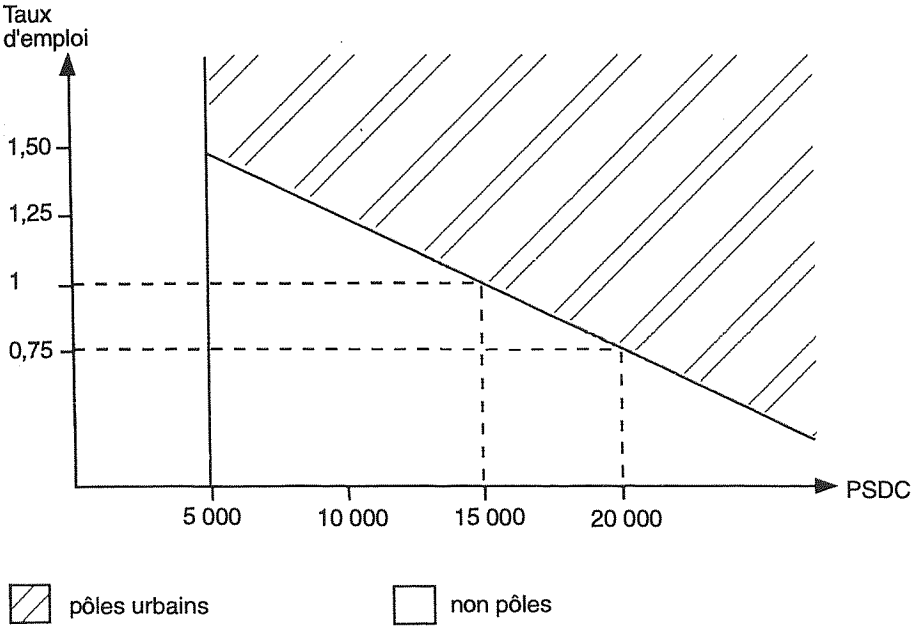
Un simple seuil de population s'avérant insuffisant, l'idée est venue de prendre en compte également le taux d'emploi, rapport des emplois au lieu de travail aux actifs résidents ayant un emploi. Ce taux d'emploi permet de prendre en compte le contexte spatial. A même niveau de population, une unité urbaine proche d'un grand pôle sera peu attractive. Elle aura un taux d'emploi faible. Ailleurs, son taux d'emploi élevé sera le signe de son rôle de ville.

La définition des pôles urbains dans le rapport intermédiaire a ainsi été de cette nature. On a éliminé toutes les unités urbaines de moins de 5 000 habitants. Et l'on a exigé aux autres unités urbaines un taux d'emploi d'autant plus élevé que leur population était plus faible. La formule était la suivante :

$$TE > 1{,}75 - PSDC/20\,000$$

où : TE = Taux d'Emploi
PSDC = Population Sans Double Compte

Cela donnait le graphique suivant :



Le taux d'emploi devait ainsi être supérieur à 1 pour une agglomération de 15 000 habitants, supérieur à 1,5 pour une agglomération de 5 000 habitants.

Dans la pratique, on excluait ainsi un certain nombre « d'unités urbaines dortoirs », et une seule unité urbaine de plus de 20 000 habitants (Savigny-le-Temple).

À un instant donné, cette définition était très intéressante. Le problème est qu'en évolution on pouvait voir disparaître des pôles, et cette disparition aurait été difficile à présenter.

Il était en effet fort possible qu'une agglomération voit sa population et ses emplois augmenter, et dans le même temps, son taux d'emploi diminuer : elle pouvait alors sortir du champ des pôles urbains. Quelques cas ont ainsi été constatés entre 1982 et 1990. Il aurait été peu aisé d'expliquer qu'une agglomération ait perdu son statut de pôle urbain au cours d'une période durant laquelle sa population et ses emplois ont augmenté.

On est donc revenu à une formulation plus simple, et donc plus facile à expliquer. Plutôt que de revenir à un seuil de population, on a choisi un seuil d'emploi, ce qui sous-entend un certain niveau d'attractivité. L'on a vérifié en outre qu'il n'y avait pas attraction du pôle considéré vers un autre pôle urbain.

La définition finalement retenue est la suivante :

Pôle urbain : unité urbaine offrant 5 000 emplois ou plus et n'appartenant pas à la couronne périurbaine d'un autre pôle urbain.

Tous les pôles urbains ainsi définis ont plus de 10 000 habitants (excepté un seul) mais la réciproque n'est pas vraie.

5. Aires urbaines - communes périurbaines

Très vite, il est apparu que l'élément déterminant dans la définition du périurbain était l'attraction qu'exerçaient les emplois de la ville sur celui-ci. Le niveau de migrations alternantes en est l' expression idéale, mais quel niveau choisir avec un minimum d'arbitraire ?

L'application du logiciel ZONAGE a d'abord été proposée, son avantage étant de ne pas décider d'un seuil a priori. Ce logiciel intègre en effet dans son fonctionnement trois contraintes non paramétrées :

1. Maximisation du nombre de zones,
2. Minimisation des échanges entre zones,
3. Cohérence du zonage, chaque commune appartenant à la zone à laquelle elle est le plus liée.

Utilisé avec le nombre de sortants comme lien, ce logiciel avait déjà permis de proposer une partition complète et unique du territoire en bassins d'emploi. Un maximum de zones cohérentes avait été obtenu, chaque commune étant rattachée à la zone où elle envoyait le plus d'actifs résidents.

Utilisé avec le nombre de sortants et de stables comme lien, le même logiciel permettait de définir également des espaces périurbains. Les communes ayant une majorité d'actifs stables, c'est-à-dire travaillant dans leur commune de résidence, constituaient autant de zones monocommunales. Il s'agissait pour l'essentiel de communes rurales. À l'inverse, les communes au lien devenu au fil du temps minoritaire avec elles-mêmes, constituaient des zones multicommunales. Parmi ces zones, celles incluant un ou plusieurs pôles urbains pouvaient être qualifiées d'aires urbaines.

L'aspect « boîte noire » de la procédure a conduit à ne pas la retenir. S'il en est fait mention ici, c'est que cette méthode a servi de point de référence pour le choix du seuil fixe de migrations altrenantes finalement retenu.

Dans un premier temps, il a été convenu d'utiliser un critère simple : le pourcentage d'actifs de chaque commune rurale ou unité urbaine allant travailler vers un pôle urbain ; ce pôle ayant été défini précédemment. Après divers tests, et au vu de la comparaison avec les résultats du logiciel ZONAGE, on a abouti à un seuil de 33,3 %, soit un tiers d'actifs.

Chaque pôle urbain était ainsi entouré d'une couronne périurbaine, où un tiers des actifs se rendaient quotidiennement dans le pôle pour y travailler. L'étendue de ces couronnes était bien plus restreinte que ne l'était celle des ZPIU, et elle semblait correspondre à ce vers quoi on voulait tendre. Là encore, le seuil d'un tiers a quelque chose d'arbitraire, propre à toute nomenclature.

C'est cette définition qui a été proposée dans le rapport intermédiaire. Si elle n'a soulevé quasiment aucune critique au sein de l'Insee, si ce n'est de façon minoritaire au sein du groupe de travail, elle a fait réagir un universitaire américain de passage en Finlande (mais connaissant très bien le concept de ZPIU, ça existe), M.Seymour Sacks. Sa critique était qu'on ne prenait en compte que l'attraction des emplois du pôle urbain lui-même. Or ce qui se rencontre fréquemment aux Etats-Unis et qui a tendance à se retrouver chez nous, c'est que les entreprises ont souvent intérêt à se délocaliser en périphérie des villes, et parfois en dehors même des limites de l'agglomération. Elles y trouvent du terrain bon marché, facilement accessible par la suite, pour y implanter leurs établissements. On reconnaît là certains hypermarchés souvent spécialisés ou certaines zones d'activité. Ces établissements participent pleinement à l'activité économique de la ville (au sens large) et doivent être pris en compte dans l'étendue de l'espace périurbain.

Pour les prendre en compte, il faut utiliser un logiciel du type MIRABELLE ou ZONAGE. Ces logiciels utilisent dans leurs agrégations successives ce que l'on nomme de manière imagée l'effet « boule de neige ». Ce type de méthode avait effectivement été envisagé lors des premières réunions du groupe, mais avait été laissé de côté pour sa relative complexité. On est finalement revenu en arrière et on a préféré pour cette fois le côté complexe mais plus satisfaisant.

Ainsi, lors de la délimitation des couronnes périurbaines, on commence dans une première étape par sélectionner les communes rurales ou unités urbaines, envoyant un certain pourcentage de migrants alternants vers le pôle urbain. Dans une deuxième étape sont ajoutées celles qui vérifient la condition par rapport à l'ensemble provisoirement formé par le pôle urbain et les communes sélectionnées à la première étape ; etc....

On aboutit, pour un même seuil de migrants alternants, à des couronnes périurbaines plus étendues. Pour limiter cette étendue, on a relevé le seuil à 40 % *(annexe 4).*

Les définitions finalement retenues sont donc les suivantes :

Aire urbaine : ensemble de communes, d'un seul tenant et sans enclave, constitué par un pôle urbain et par des communes rurales ou unités urbaines dont au moins 40 % de la population résidente ayant un emploi travaille dans le pôle ou dans des communes attirées par celui-ci.

6. La multipolarité

La plus grande originalité dans les conclusions du groupe tient à cette volonté de prendre en compte le phénomène de multipolarité. Très rapidement dans les discussions est venue l'idée qu'une commune fortement attirée par des emplois urbains, situés dans différents pôles, était également périurbaine.

Il n'était pas question de faire du typologique dans cette partie du zonage. Au niveau de migrations alternantes s'est donc ajoutée la contrainte de connexité : les communes, attirées par plusieurs pôles, devaient en outre former un ensemble connexe avec eux et leur couronne périurbaine. Pour que cet ensemble de communes soit un complément des aires urbaines, on a considéré les flux allant vers les aires urbaines (pôles et couronnes correspondantes).

Le critère de contiguïté et de connexité est important dans cette définition. Le simple contact entre deux communes peut créer un effet de chaîne et faire se rejoindre entre elles de nombreuses aires urbaines. Cela donne parfois un aspect plus fragile à cet ensemble connexe qu'on a dénommé espace urbain. Pour ne pas ajouter à cet effet en chaîne, on n'a pas appliqué cette fois d'effet « boule de neige ».

Pour rester cohérent avec la définition des aires urbaines, on a pris le même seuil de migrants alternants : 40 %. Les communes ainsi définies ont pris le nom de communes multipolarisées *(annexe 5)*.

L'algorithme de délimitation des espaces urbains opère de façon descendante :

1. On repère tous les atomes (communes rurales ou unités urbaines non pôles) dont plus de 40 % des actifs travaillent dans l'ensemble des aires urbaines de France métropolitaine. On établit la liste des zones connexes Z1 ainsi formées par ces communes, plus les aires urbaines.

2. Parmi l'ensemble des atomes d'une zone Z1, on repère ceux dont plus de 40 % des actifs travaillent dans les aires urbaines de cette zone. On établit une nouvelle liste de zones connexes Z2.

3. On itère - les zones étant décroissantes et incluant les aires urbaines - il y a nécessairement convergence. Le processus converge pour le RP90 en 3 étapes

Z1 à Z3. La carte Z3 est celle des espaces urbains. On remarque qu'apparaissent d'assez nombreuses zones multipolaires - ce qui était recherché par cette méthode.

44 espaces urbains multipolaires ont ainsi été définis en 1990. Celui contenant Paris est le plus important ; il comprend 44 aires urbaines.

Pour éclairer la hiérarchie et les liaisons entre les différentes aires qui constituent chaque espace multipolaire, on a à nouveau utilisé le logiciel MIRABELLE. Il a été appliqué pour chacun d'entre eux à partir des éléments géographiques suivants : chaque aire urbaine prise dans sa globalité, et chaque commune multipolarisée. On obtient ainsi un graphe qui permet de visualiser les hiérarchies et les emboîtements entre les différentes aires.

La définition de ces communes multipolarisées et des espaces urbains vient en dernier pour compléter l'espace à dominante urbaine. Mais le choix de mettre en lumière cette multipolarité n'a pas été sans conséquence sur la définition des aires urbaines. En l'absence de cette notion de communes multipolarisées à 40 %, on aurait pu éventuellement définir plusieurs couronnes périurbaines autour du pôle urbain, avec différents seuils, 40 %, 30 %, 20 %.

Les définitions retenues sont les suivantes :

. **Communes multipolarisées** : communes rurales et unités urbaines situées hors des aires urbaines, dont au moins 40 % de la population résidente ayant un emploi travaille dans plusieurs aires urbaines, sans atteindre ce seuil avec une seule d'entre elles, et qui forment avec elles un ensemble d'un seul tenant.

. **Espace urbain multipolaire** : ensemble d'un seul tenant de plusieurs aires urbaines et des communes multipolarisées qui s'y rattachent.

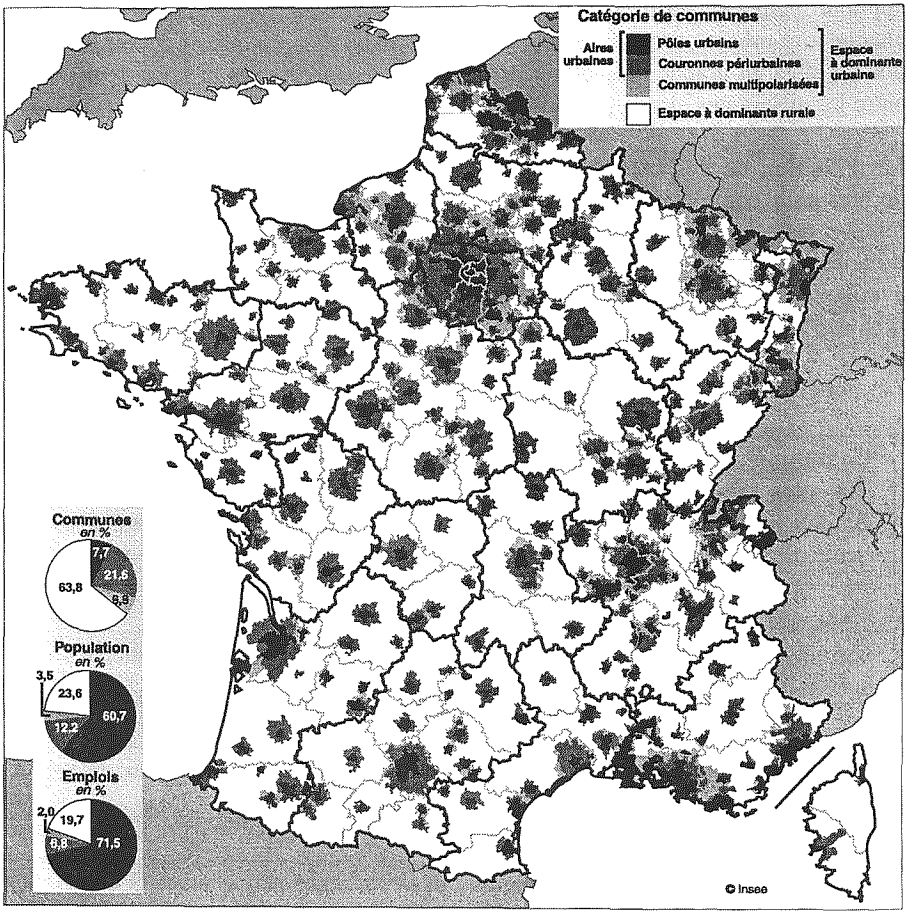
Espace urbain monopolaire : aire urbaine qui ne fait partie d'aucun espace urbain multipolaire.

. **Communes périurbaines** : communes des couronnes périurbaines et communes multipolarisées.

. **Espace à dominante urbaine** : ensemble des pôles urbains et des communes périurbaines ; peut se définir également comme l'ensemble des aires urbaines et des communes multipolarisées ; ou encore comme l'ensemble des espaces urbains, monopolaires et multipolaires.

On a représenté sur la carte 2 les 36570 communes de la France métropolitaine suivant leur appartenance aux différents types.

Carte 2-
Le zonage en urbaines 1990
(ZAU)



7. L'espace à dominante rurale

Cet espace a finalement été défini comme suit :

. **Espace à dominante rurale** : ensemble des communes rurales et unités urbaines n'appartenant pas à l'espace à dominante urbaine.

Cet espace demeure défini de manière résiduelle par rapport à l'espace à dominante urbaine.

Il n'a pas été décomposé, comme cela était prévu aux débuts du groupe de travail, et comme cela était demandé par de nombreux interlocuteurs.

Le groupe a, en cours de route, revu à la baisse ses ambitions et décidé de ne s'occuper finalement que de l'espace à dominante urbaine. L'espace restant, dit à dominante rurale, comprend des communes rurales faiblement attirées par les pôles urbains, mais également des petites unités urbaines n'ayant pas assez d'emplois pour être qualifiées de pôles urbains, mais suffisamment sans doute pour avoir leur propre autonomie. On voit bien que cet espace est loin d'être homogène, entre des communes faiblement sous influence urbaine, mais peut-être pas de manière négligeable, des unités urbaines jouant localement un certain rôle, et des communes rurales éloignées de tout.

Dans un rapport intermédiaire, le groupe de travail proposait une décomposition plus fine de cet espace. Prendre un seuil plus bas pour les migrations alternantes a été testé (10% de migrations alternantes) mais c'était trop faible pour être pertinent.

De nombreux spécialistes du monde rural considèrent de toutes façons que cet espace ne se structure pas de la même manière que l'espace à dominante urbaine. En revanche ce qui est primordial, c'est le niveau d'équipement et le rayonnement qu'il procure aux petits pôles bien dotés, les bourgs centre.

Au préalable, le groupe de travail a passé en revue un certain nombre de critères qui ont pu être évoqués ici ou là :

- la surface agricole utile (SAU) ne parait plus être discriminante dans la mesure où l'agriculture ne joue plus un rôle prépondérant en terme d'actifs. De plus, elle fournit une description de la manière dont est utilisé le territoire pour la production agricole, et ne renseigne pas sur l'occupation du territoire par la population (l'Ile-de-France par exemple reste une grande région agricole).

- la proportion de résidences secondaires qui peut être le signe d'une pré-urbanisation, ou d'un dynamisme touristique, n'a pas été retenue du fait de sa médiocre qualité statistique.

- la part de réserves foncières provient d'une source difficile à mobiliser.

- le poids des inactifs ou des personnes âgées a été jugé d'utilisation difficile.

- la part de construction neuve a été également évoquée sans être retenue.

- enfin la densité de population, fréquemment citée, a été jugée d'une utilisation délicate. Le maillage communal n'est pas du tout homogène sur l'ensemble du territoire, et les lissages du genre « températures urbaines » ne semblent pas lever complètement la difficulté.

L'option a donc été prise de travailler à partir des informations de l'inventaire communal, qui, à l'époque, devait précéder d'une année le recensement de la population.

À partir de l'inventaire communal ont été sélectionnés 22 équipements. Cette liste d'équipements a été obtenue à partir d'une classification hiérarchique ascendante sur les indices de co-attractivité de 55 équipements (cf. Insee-Première numéro 325 - juin 1994 : « L'influence des villes et des bourgs-centres » V.Vallès-P Hugon). Les équipements caractéristiques des bourgs-centres sont les commerces non alimentaires, les services financiers (banque, caisse d'épargne, étude de notaire), les services locaux de l'Etat (gendarmerie, perception, collège) et certains professionnels de santé tels : le dentiste, le kinésithérapeute, le vétérinaire et l'ambulancier.

On a alors défini des pôles de services qui possédaient au moins 16 équipements sur la liste des 22.

Ensuite a été défini un espace rural proche comprenant toutes les unités urbaines non encore sélectionnées jusqu'à présent, ainsi que toutes les communes rurales n'appartenant pas à une aire urbaine, n'étant pas pôle de services, mais dont la distance moyenne aux 22 équipements est inférieure à 12 km.

Pour chaque équipement, la distance est calculée par rapport à la commune où l'équipement est réellement fréquenté (tel que cela apparaît dans l'inventaire communal).

Enfin, l'espace rural éloigné comprenait les communes rurales dont la distance moyenne aux 22 équipements est supérieure à 12 km.

Cette décomposition de l'espace à dominante rurale, présentée dans le premier rapport, n'a malheureusement pas été validée par les directions régionales. Cela est dû sans doute en partie à cause de l'inventaire communal lui-même qui donnait des informations par commune parfois incohérentes entre elles. Cela était dû plus sûrement à cette notion de distance calculée à vol d'oiseau qui n'a pas la même signification d'une région à l'autre. Certaines régions ne retrouvaient pas ce qu'elles croyaient être leur rural profond.

D'autres trouvaient du rural éloigné tout près des centres urbains.

Il a donc été décidé de laisser de côté pour l'instant cette décomposition. On peut signaler cependant que la porte reste ouverte. Elle a d'ailleurs été franchie pour une prochaine publication sur « les espaces ruraux » dans la collection Contours et Caractères . Une décomposition plus fine de l'espace à dominante rurale était indispensable pour la réalisation de cet ouvrage. Elle a été réalisée par un petit groupe Inra/Insee tout en respectant l'espace à dominante urbaine.

Cette décomposition n'a toutefois pas le caractère de nomenclature Insee comme le zonage en aires urbaines.

Elle a été bâtie sur le critère des migrations alternantes, une valeur intermédiaire de 20 % étant retenue.

Ont donc été distinguées au sein de l'espace à dominante rurale : des communes qui, sans se trouver sous une forte dépendance de la ville, sont néanmoins sous une influence urbaine plus diffuse ; des communes ou unités urbaines qui, du fait d'un nombre d'emplois encore conséquent et d'une certaine attractivité vis-à-vis de leur environnement, peuvent être considérées comme de petits pôles d'emplois ; des communes placées sous l'influence de ces petits pôles ; enfin les autres communes forment alors une catégorie de rural isolé. Les communes appartenant à l'espace à dominante rurale sont ainsi réparties selon les quatre catégories suivantes :

- du rural sous faible influence urbaine : il s'agit des communes rurales ou unités urbaines dont 20 % ou plus des actifs vont travailler dans l'une quelconque des aires urbaines définies dans le ZAU (*carte 3*) ;

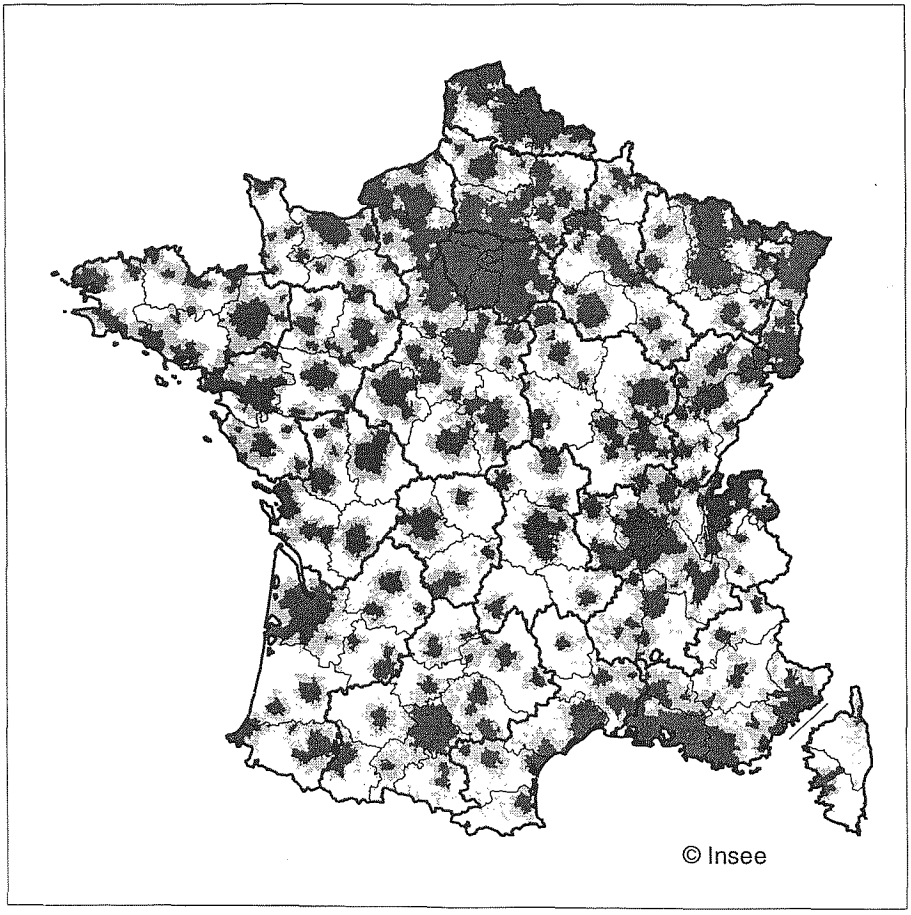
- des pôles ruraux : sont considérées comme telles les communes rurales ou unités urbaines qui regroupent 2 000 emplois ou plus (c'est-à-dire entre 2 000 et 5 000 emplois) et dont le nombre d'emploi est supérieur ou égal au nombre d'actifs résidents (taux d'emploi >= 1) (*carte 4*) ;

- du rural sous l'influence des pôles ruraux : il s'agit ici des communes rurales ou unités urbaines dont plus de 20 % des actifs résidents vont travailler dans l'un quelconque des pôles ruraux. Il ne nous a pas semblé nécessaire de distinguer, à

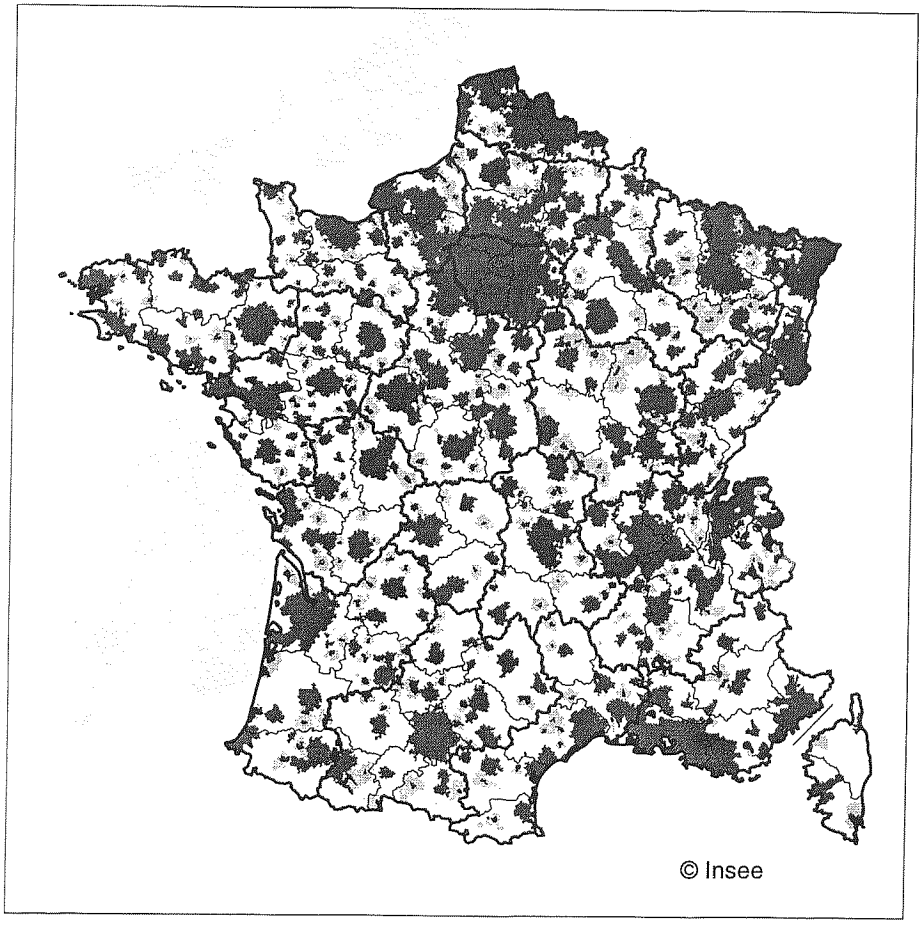
l'instar de ce qui a été fait pour les pôles urbains, les sphères respectives de ces petits pôles. Du fait du caractère géographiquement limité de leur influence, on a préféré repérer globalement celle-ci.

- du rural isolé : celui-ci est constitué de toutes les autres communes rurales ou unités urbaines (*carte 5*).

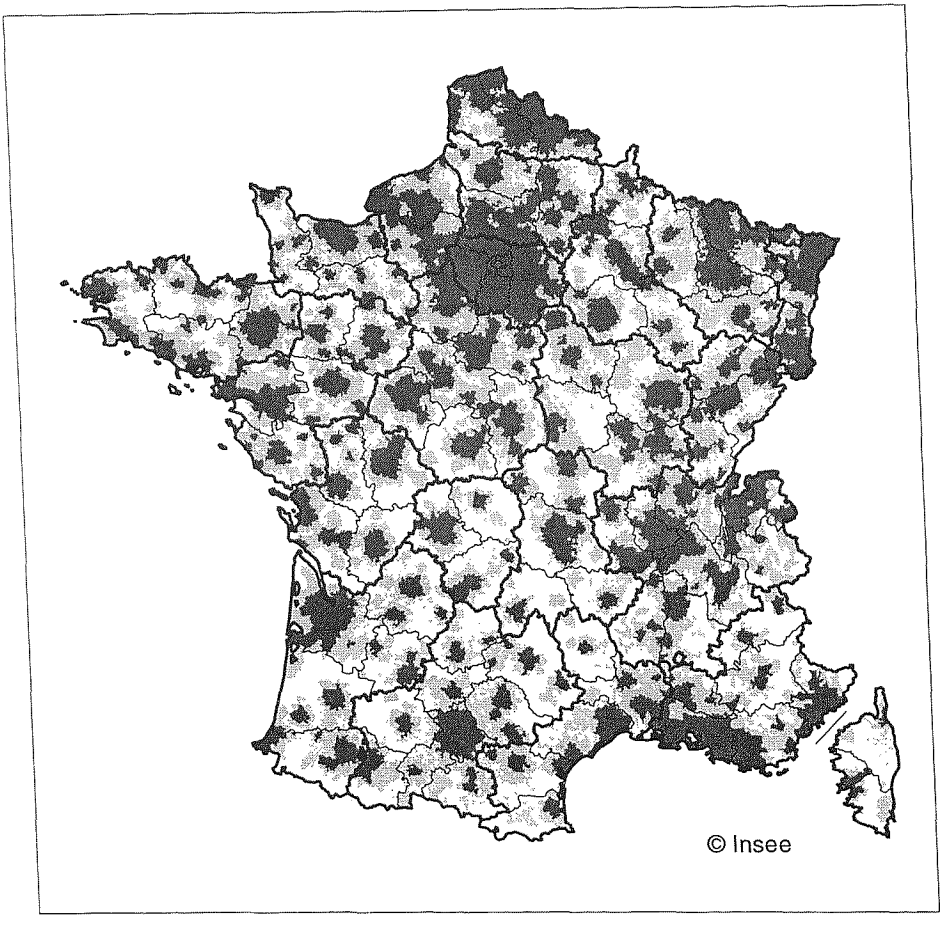
Carte 3-
Communes sous faible influence des aires urbaines



Carte 4-
Pôles ruraux et leur aire d'influence



Carte 5-
Rural isolé



Catégorie de communes

- Espace à dominante urbaine
- Pôles ruraux et leur aire d'influence et communes sous faible influence des aires urbaines
- Rural isolé

Bibliographie

Camus M., Ricci C. (1996), « Aires urbaines : au-delà des agglomérations », Insee Bourgogne, *Dimensions* n° 36, août.

Heller J.L., Laganier J. (1996), « 4 habitants sur 5 de la façade méditerranéenne résident dans les espaces urbains », Insee Languedoc-Roussillon, *Repères pour l'économie Languedoc-Roussillon,* n° 22, 3ème trimestre.

Insee (1990), Composition communale des unités urbaines : population et délimitation 1990.

Insee (1992), Composition communale des zones de peuplement industriel ou urbain : population et délimitation 1990.

Insee (1997), Composition communale du zonage en aires urbaines : population et délimitation 1990, *Nomenclatures et codes,* à paraître.

Insee Bretagne (1996), « Les aires urbaines », *Octant,* n° 65, juin.

Insee-Inra (1997), *Les espaces ruraux*, Insee, Contours et caractères, à paraître.

Jacquier J. (1996), « Nouveaux espaces », Insee Pays de la Loire, *Référence Pays de la Loire,* n° 14, avril.

Julien P. (1996), « Urbain et rural : une nouvelle approche, « Insee-Midi-Pyrénées *6 pages*, n° 10, février.

Laganier J., Le Jeannic T. (1996), Rapport intermédiaire : « Aires d'attraction urbaine et espace rural pour succéder aux ZPIU/hors ZPIU », note interne Insee du 28 octobre 1994 n° 246/H323/TLJ-JL/DP.

Laganier J., Le Jeannic T. (1996), « Rapport final pour le remplacement des ZPIU », note interne Insee du 12 mars 1996 n° 48/H323/TLJ/DP.

Le Gléau J.P., Pumain D., Saint-Julien T. (1996), « Villes d'Europe : à chaque pays sa définition », *Économie et statistique,* n° 294-295.

Le Jeannic T. (1996), « Une nouvelle approche territoriale de la ville », *Économie et statistique,* n° 294-295.

Le Jeannic T. (1996), « Migrations et croissance urbaine », in *Données urbaines,* sous la direction de D.Pumain et F.Godard, PIR-Villes, Anthropos éditeur.

Le Jeannic T. (1997), « Pôles urbains et périurbanisation : le zonage en aires urbaines », *Insee première*, n° 516, 1997.

Le Jeannic T. (1997), « Trente ans de périurbanisation », *Économie et statistique,* à paraître.

Loonis Vincent (1996), « Le ZAU remplace les ZPIU », *Insee Aquitaine* n° 37, juillet.

Marpsat M. (1993), « Déchiffrer la ville », *Courrier des statistiques,* n° 67-68, décembre, pp. 27-35.

Robert I. (1996), « Urbain, rural : des concepts qui évoluent », Insee Centre, *L'économie du Centre* n° 15, octobre.

Técher G. (1996), « Le zonage en aires urbaines », Insee revue *Champagne-Ardenne* n° 5

Vallès V. (1996), « les aires urbaines », Insee Auvergne, *Le point économique de l'Auvergne* » n° 47, juillet 1996.

Willm Y., Court Y., Gauthier O. (1996), « Les espaces urbains, l'espace rural, une nouvelle partition du territoire », Insee Poitou-Charentes, *Décimal* n° 165, avril.

ANNEXE 1

LISTE DES PERSONNES RENCONTRÉES

Personnes internes à l'INSEE :

M. DE LOS SANTOS	Chef de la Division Agriculture
M. HILAL	Division Agriculture
Mme MARPSAT	Chef de la Division Études Sociales
M. TERRIER	Mission Systèmes d'Études et de Diffusion de Données Locales
M. ANFRÉ	Chef de la Division Statistiques Communales et Locales
M. COURSON	Division Statistiques Communales et Locales
M. DESPLANQUES	Chef de la Division Études et Enquêtes Démographiques
M. CASTELLAN	Mission Ville
M. LE FILLATRE	Direction des Statistiques Economiques
M. RONSAC	Direction Régionale d'Île-de-France
M. LE JEANNIC	Direction Régionale d'Île-de-France
M. JULIEN	Direction Régionale de Midi-Pyrénées
M. LAURENT	Direction Régionale de Bretagne

Personnes extérieures

M. JAYET	Université de Lille

Mme SAINT-JULIEN	Centre National de Recherche Scientifique
M. SCHMITT	Institut National de la Recherche Agronomique
M. CAVAILHES	Institut national de la Recherche Agronomique
M. BIRABEN	Institut National des Études Démographiques
M. BONTRON	SEGESA
M. KAYSER	Université du Mirail Toulouse
M. MENDRAS	Observatoire Français des Conjonctures Economiques
M. LE BRAS	École des Hautes Études en Sciences Sociales
Mme RATTIN	Service Statistique du Ministère de l'Agriculture
Mme CAVALIER	Service Statistique du Ministère de l'Agriculture
M. LUSSON	Fédération Nationale des Agences d'Urbanisme
MM. DUPORT, LERY MM. GASTAMBIDE, PHILIZOT M. WELLHOFF, Mme DELAMARRE Mme HAUTROUCHARD	Délégation à l'Aménagement du Territoire et à l'Action Régionale

MEMBRES DU GROUPE DE TRAVAIL

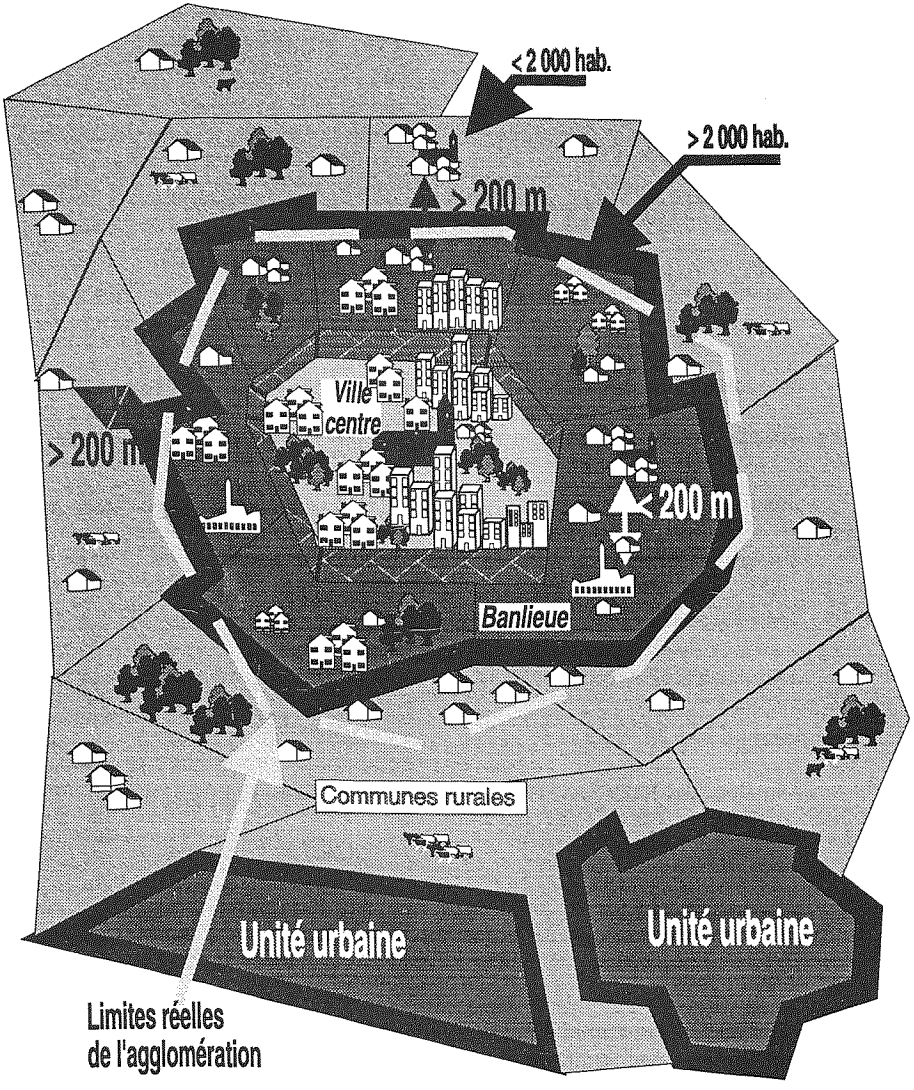
Co-parrainé par les Division « Recensements » et « Statistiques et Études Régionales », ce groupe de travail était composé des membres suivants :

Marc CAMUS	DR de Bourgogne
Jean-Marie DUVAL	DR d'Aquitaine
Jean-Christophe FANOUILLET	DG/DSDS
Rémy FERRON, Daniel GUILLEMET	DR de Bretagne
Jean-Luc HELLER	DR de Languedoc-Roussillon
Mohamed HILAL	DG/DSE
Philippe JULIEN	DR de Midi-Pyrénées
Jean LAGANIER	DR de Provence -Alpes-Côte d'Azur
Loeiz LAURENT	DG/DDAR
Jean-Pierre LE GLÉAU	DG/DCSRI
Thomas LE JEANNIC	DG/DDAR
Chantal MADINIER	DG/DSDS
Jean-Jacques RONSAC *(*)*	DR Ile-de-France
Christophe TERRIER	DG/DDAR
Vincent VALLÈS	DR d'Auvergne

(*) Par son extraordinaire connaissance du territoire français et son expérience sur les zonages, Jean-Jacques RONSAC a été d'une très grande aide dans nos réflexions. Il nous a malheureusement quitté en Août 1994. Les membres du groupe de travail se souviendront de lui comme d'une personne particulièrement chaleureuse et généreuse, qui savait communiquer sa passion pour son travail

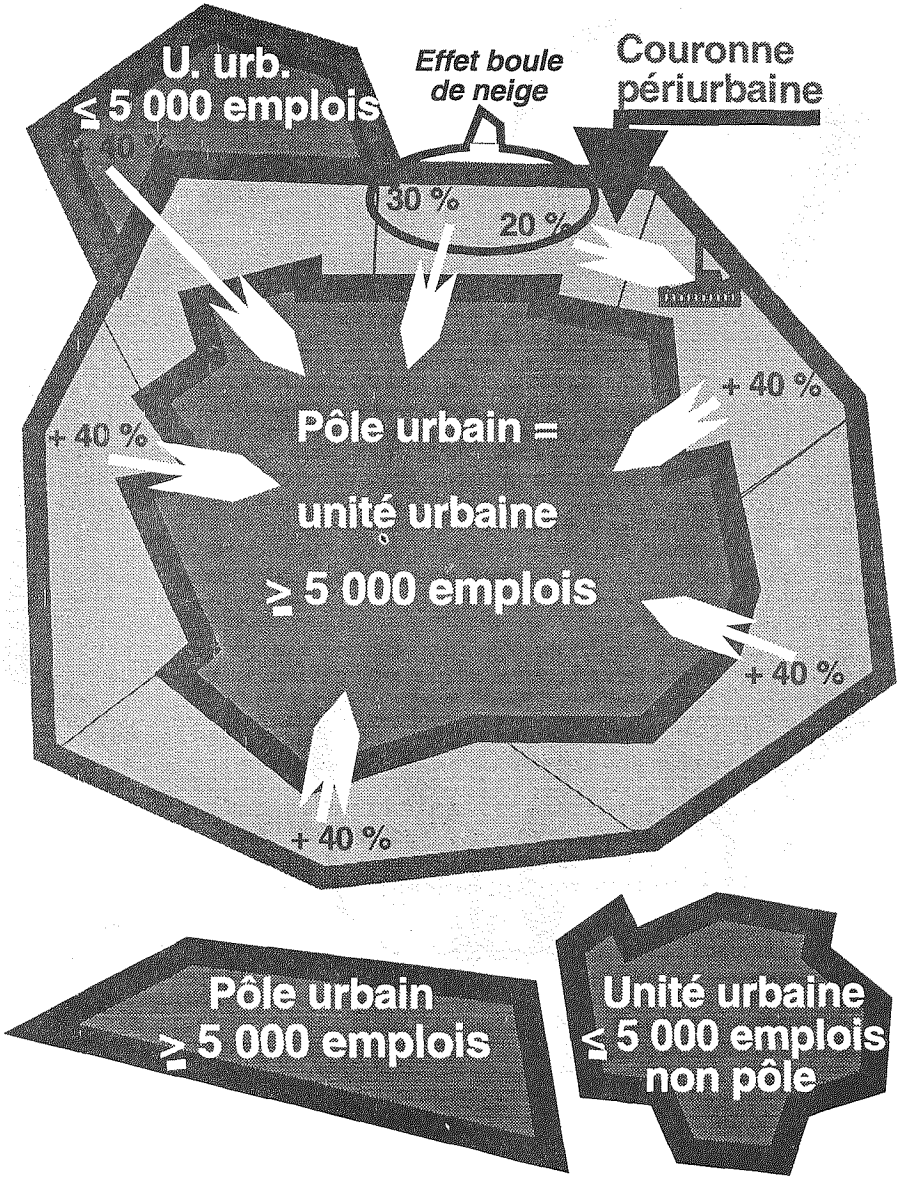
Unité urbaine (ville centre + banlieue)

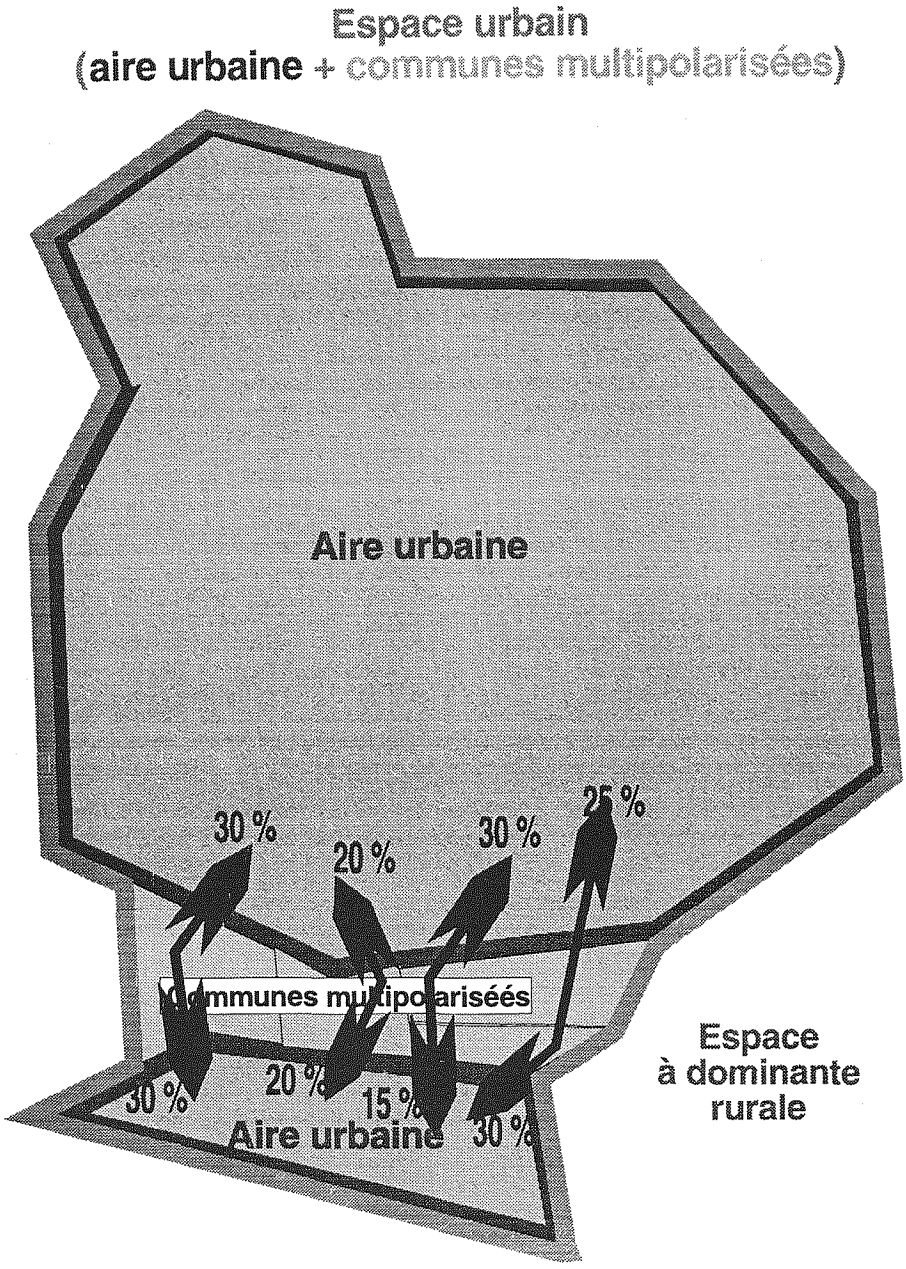
- agglomération de **plus de 2 000 habitants**
- agglomération : zone bâtie avec **moins de 200 m** entre les habitations



Aire urbaine (pôle urbain + couronne périurbaine)

Pôle urbain = Unité urbaine de + de 5 000 emplois




Espace urbain
(aire urbaine + communes multipolarisées)
Aire urbaine
30 %
20 %
30 %
25 %
Communes multipolarisées
30 %
20 %
15 %
Aire urbaine
30 %
Espace
à dominante
rurale

Session 5

Les enquêtes sur des sujets sensibles

BILAN MÉTHODOLOGIQUE DES ENQUÊTES FRANÇAISES SUR LE COMPORTEMENT SEXUEL

Alain Giami, Alfred Spira (Inserm) et les groupes ACSF, ACSAG et ACSJ

À la fin des années 80, l'Agence Nationale de Recherches sur le Sida (ANRS) a confié à l'INSERM la responsabilité de trois enquêtes relatives aux comportements sexuels : l'enquête ACSF auprès des adultes en métropole, ACSAG auprès des adultes en Guadeloupe, Martinique et Guyane, ACSJ auprès des jeunes de 15 à 18 ans en métropole. Les résultats de ces enquêtes ont fait l'objet de nombreuses publications et le rapport scientifique et méthodologique de l'enquête ACSF sera très prochainement publié.

L'objectif de ces trois enquêtes est double : apporter une connaissance précise des comportements sexuels et éclairer les politiques de prévention de l'épidémie de Sida. Il s'agit de décrire, d'une part, les comportements sexuels et la dynamique personnelle des désirs et des attitudes, et d'autre part, l'appropriation d'une logique de prévention. L'enquête a en outre pris en compte les biographies individuelles, les relations entre partenaires, les normes et les représentations sociales de la sexualité et l'influence des réseaux de confidents.

Il fut de plus décidé, en 1994, de réaliser une enquête plus légère (Moatti al., 1995) visant à mesurer, l'évolution des connaissances et attitudes vis-à-vis de l'infection, et celle de quelques indicateurs de comportements sexuels et de protection contre le VIH et les MST. Le problème de la pérennisation d'un tel instrument de mesure est maintenant à l'ordre du jour. Celui-ci doit être conçu comme un instrument de surveillance et non de recherche à proprement parler.

Les difficultés soulevées par de telles enquêtes sont exemplaires : comment interroger quelqu'un sur sa sexualité, en respectant sa sensibilité ? Quels mots employer ou ne pas employer ? Quelle langue utiliser, notamment aux Antilles ou en Guyane, le français ou le créole ? Quels questionnements particuliers dans ces départements d'outre-mer ? Quel sens donner aux divergences des réponses des hommes et des femmes concernant leurs nombres de partenaires ? Quelles sont les logiques de comportement face au risque du sida ? Une telle recherche a nécessité le travail en commun d'une équipe pluri-disciplinaire composée d'épidémiologistes, psycho-sociologues, linguistes, sociologues, économistes, démographes et statisticiens !

Un travail d'une telle envergure était d'autant plus nécessaire qu'aucune enquête quantitative sur les comportements sexuels n'avait été menée en France, en population générale, depuis le rapport Simon (1972) et que ce dernier ne considérait pas les DOM : centré explicitement sur les comportements sexuels, le rapport Simon visait à comprendre les facteurs psycho-sociaux qui favorisent l'utilisation de la contraception ainsi que les modifications de la sexualité et des rapports entre les sexes suscités par cette utilisation. Par ailleurs, les premières enquêtes de l'INED sur la contraception en 1971 et 1978 abordaient déjà certaines pratiques sexuelles de manière allusive (Collomb et Zucker, 1977) et (Léridon et al. 1979).

Dans une discussion des principaux problèmes méthodologiques rencontrés dans les enquêtes quantitatives sur les comportements sexuels à l'époque du sida, J. Catania et ses collègues (1995) considèrent que de nombreux chercheurs ont adopté une attitude conservatrice frileuse par rapport à l'innovation méthodologique en n'essayant pas de bénéficier des innovations techniques qui permettent d'améliorer la qualité des données d'enquêtes. Dans certaines situations nationales, où les chercheurs ont rencontré de nombreux obstacles politiques à la réalisation des enquêtes, le choix a été fait d'adopter les méthodes de recherche les plus classiques et les instruments les plus validés (Laumann, Gagnon, Michael, Michaels, 1994). L'élaboration des instruments de recherche et l'innovation méthodologique relèvent donc de choix théoriques aussi bien que stratégiques et politiques.

Les enquêtes françaises ont été réalisées dans un contexte politique différent qui a incité les chercheurs à élaborer une double problématique : sexualité et sida. Les méthodologies ont été bien différenciées selon le public enquêté : après expérimentation, l'enquête ACSF a été réalisée par téléphone, tandis que l'enquête ACSAG était réalisée en face à face au domicile des adultes enquêtés et l'enquête ACSJ en face à face dans les établissements scolaires ou d'apprentissage.

Qui interroger ?

Population générale, groupes à risque et comportements à risque

La définition de la population à étudier constituait l'un des problèmes majeurs à résoudre pour l'enquête ACSF. Il fallait mettre en œuvre une enquête permettant de produire des connaissances sur les principaux comportements sexuels de la population générale et en même temps apporter des informations sur les "groupes à risque". Or en France comme dans d'autres pays, dans le contexte du Sida, la définition des "groupes à risque" constitue un enjeu politique au moins autant que scientifique (Réseau National de Santé Publique, 1996). Doit-on considérer que seuls les comportements et les attitudes des personnes et des groupes considérés a

priori comme " à risque " doivent être étudiés dans une perspective limitée, ou bien que la compréhension globale de la dynamique de l'épidémie, ainsi que sa prévision, passe par une analyse des principaux comportements sexuels et des attitudes de l'ensemble de la population générale, permettant d'évaluer les risques potentiels encourus par celle-ci et par les différents sous-groupes qui la composent ? Dans le cadre d'une enquête en population générale, l'accès à des groupes minoritaires est loin d'être facile, a fortiori lorsqu'il s'agit de groupes dont les membres ont des comportements stigmatisés (homosexualité, usage de drogues par voie intraveineuse, notamment). La difficulté d'accéder aux individus appartenant à des groupes minoritaires ne doit pas masquer le problème qui consiste à évaluer la place relative de ces groupes dans l'ensemble de la population générale, particulièrement lorsqu'il s'agit de comprendre la dynamique d'une épidémie en partie liée à des comportements. Doit-on considérer que seuls les comportements et les attitudes des personnes et des groupes considérés a priori comme " à risque " doivent être étudiés dans une perspective limitée ou bien que la compréhension globale de la dynamique de l'épidémie passe par une analyse des comportements sexuels et des attitudes de l'ensemble de la population générale permettant d'évaluer les risques potentiels encourus par celle-ci et par les différents sous-groupes qui la composent ? Le dispositif ACSF et notamment l'élaboration et d'un plan de sondage complexe, résultent de compromis entre ces deux objectifs théoriques.

Échantillon représentatif et inférence probabiliste

Le problème de l'échantillonnage a traditionnellement reçu des solutions différentes selon les disciplines. Les chercheurs en sciences sociales se sont très tôt préoccupés de construire des échantillons de populations considérés comme représentatifs de l'ensemble de la population et permettant une extrapolation à celle-ci. L'épidémiologie analytique, discipline quantitative qui vise à l'identification des facteurs de survenue et d'évolution des maladies, repose, pour sa part, sur la constitution d'échantillons qui ont pour but la comparabilité plus que la représentativité. Devant la difficulté de constituer des échantillons de personnes représentatives des malades, pouvant être comparés à la population générale, l'une des méthodes les plus employées repose sur la constitution d'échantillons témoins, permettant des analyses comparatives par rapport à des échantillons de malades (Bouyer J et al. 1993). Les problèmes d'inférence sont alors résolus d'un point de vue probabiliste. Ce n'est que récemment que les développements de l'épidémiologie descriptive ont amené les chercheurs de cette discipline à se préoccuper de plus en plus de la constitution d'échantillons représentatifs de la population générale, ou de certains sous-groupes de la population, particulièrement intéressants du fait de leurs caractéristiques ou de leur exposition à certains facteurs de risque. La rencontre, lors de la réalisation de l'enquête ACSF, de chercheurs issus de ces différentes disciplines, donna ainsi l'occasion de construire un dispositif

d'enquête original et complexe fondé sur l'association entre une méthode d'inférence probabiliste et une méthode fondée sur la représentativité de l'échantillon.

Taille de l'échantillon ACSF

Si l'on désirait obtenir des données suffisamment précises pour des sous-groupes de la population très minoritaires, il fut calculé que le nombre de sujets nécessaire était au total de l'ordre de 20 000 personnes. C'est d'ailleurs exactement au même ordre de grandeur qu'étaient parvenues les équipes britannique et américaine : la taille d'un échantillon dépend de la précision recherchée et de la fiabilité des analyses statistiques et non de la taille de la population dont l'échantillon est extrait.

Élaboration d'un sondage complexe

La "carte filtre" : échantillon à risque et échantillon témoin

Compte tenu du nombre important de sujets à interroger et du grand nombre de questions que l'on serait amené à poser à chacun d'entre eux, un problème " logistique " se posait. Il fut résolu de la façon suivante. Le questionnaire fut réalisé sous deux formats. Après une trentaine de questions générales, permettant de décrire le sujet interrogé, une " carte filtre "[1] permet d'orienter les sujets soit vers un questionnaire court (administré à 15 235 personnes), abordant de manière schématique les principaux éléments de la vie sexuelle des personnes qui ne présentent aucun facteur de risque particulier vis-à-vis de l'infection par le VIH, ou vers un questionnaire long, beaucoup plus détaillé. Ce questionnaire long a été administré aux personnes qui ont répondu positivement à au moins une des questions de la carte filtre. Grâce à cette tactique, le questionnaire long fut effectivement administré à 4820 personnes dont 2332 qui ont finalement été caractérisées comme non à risque du fait de leurs réponses et qui représentent l'échantillon témoin et 2187 caractérisées comme "à risque". Cette solution permit de surreprésenter les individus considérés comme potentiellement à risque et de pouvoir ainsi les comparer avec l'échantillon représentant la population générale. Ce

1 Carte filtre du questionnaire ACSF :
- Vous êtes né le 4, le 17 ou le 20 d'un mois de l'année.
- Vous avez eu des rapports sexuels avec au moins 2 personnes différentes dans les 12 derniers mois.
- Au cours des 5 dernières années, vous avez eu des rapports sexuels au moins 1 fois avec 1 personne du même sexe que vous.
- Au cours des 5 dernières années, vous avez payé au moins 1 fois pour avoir des rapports sexuels.
- Au cours des 12 derniers mois, vous avez utilisé au moins une fois une drogue douce (hasch, marijuana...) ou dure (cocaïne, héroïne...).
- Vous êtes hémophile.

plan de sondage très particulier compliqua quelque peu l'analyse statistique en imposant le recours à une méthode particulière de pondération et de calcul de la variance des estimations (Leridon, in Rapport ACSF 1993 ; Warszawski, et al., 1996).

Les "modules"

Par ailleurs, de façon à pouvoir inclure un nombre suffisant de questions d'ordre psychologique et sociologique dans le questionnaire long, l'échantillon final fut divisé en deux sous-groupes, chacun d'entre eux n'étant interrogé que pour un nombre limité de " modules ". Ainsi la moitié de l'échantillon du questionnaire long fut interrogée sur leurs " confidents ", l'autre moitié sur les " fantasmes ". Ce qui eut pour effet d'augmenter la complexité du plan de sondage.

La décision d'interroger, dans la même enquête, un échantillon " témoin " considéré comme non exposé au risque et un échantillon de personnes considéré comme potentiellement à risque d'infection au VIH constitue l'une des originalités de l'enquête ACSF. Elle a permis en particulier de repérer les facteurs qui sont corrélés aux situations de prise de risque, d'identifier les groupes considérés comme à risque et d'évaluer leur place dans la population générale.

Échantillon de l'enquête ACSAG

L'enquête ACSAG repose sur un plan de sondage plus simple, du fait de la situation épidémiologique particulière des Départements Français d'Amérique (DFA) et de la taille relativement petite de la population des Antilles et de la Guyane qui ne permet pas une enquête de l'ampleur de celle réalisée en Métropole. La taille des échantillons a donc été définie à partir de la seule proportion estimée de multipartenaires dans la population. En effet, selon les données épidémiologiques, les rapports hétérosexuels sont de loin la première cause de contamination dans ces pays.

L'équipe ACSAG fit ainsi l'économie du questionnaire court de l'enquête ACSF, son questionnaire étant, en effet, une adaptation aux réalités des départements d'outre-mer de la version longue d'ACSF.

La difficulté principale du sondage tenait à la très forte hétérogénéité des structures géo-démographiques des Antilles et de la Guyane française. Par exemple, en Guyane, toutes les communes longeant les fleuves Maroni et Oyapok ont été exclues du champ de l'enquête. Les spécificités socioculturelles des groupes des Noirs réfugiés et des Amérindiens qui résident en forte proportion dans ces communes rendaient, en effet, inadaptées les procédures d'une enquête quantitative sur la sexualité dans les habitats traditionnels de ces groupes.

Il a donc été choisi de constituer l'échantillon de l'enquête, dans chaque département, par la méthode du tirage aléatoire. Les échantillons ont été stratifiés par commune afin d'assurer une bonne représentativité de la répartition de la population de chaque département.

En Martinique comme en Guadeloupe, la liste des abonnés au téléphone a pu être retenue comme base de sondage. En Guyane, la faiblesse de la couverture téléphonique a conduit à tirer au sort les logements à enquêter à partir de photographies aériennes et de plans de cadastre.

Les individus sélectionnés devaient être âgés de 18 à 69 ans, parler le français ou le créole et avoir leur résidence principale sur le lieu où se déroulait l'enquête. Au total 1007 questionnaires ont été collectés en face à face en Guadeloupe, 1006 en Martinique et 621 en Guyane.

Échantillon de l'enquête ACSJ

L'enquête ACSJ, menée au premier trimestre 1994 auprès de 6500 jeunes âgés de 15 à 18 ans, nécessitait une approche spécifique. Contrairement à l'enquête ACSF, réalisée par téléphone auprès des adultes, cette enquête a été menée entièrement en face à face dans les lycées et centres de formation des apprentis (CFA, CIPA). Cet aspect institutionnel est important à double titre : l'Education Nationale est une institution sensible au sein de laquelle il est difficile de mener des opérations si délicates ne peuvent être menées qu'avec une précaution extrême, mais d'autre part sa coopération active fut un atout précieux pour le sondage et la collecte.

Après classement des départements selon la prévalence du sida, 18 d'entre eux ont été retenus. Des établissements, stratifiés par taille, ont ensuite été tirés : 116 lycées, 50 CFA, 30 CIPA et 25 organismes de formation ont été enquêtés. La plus lourde tâche du sondage a consisté à établir la liste des centres de formation des apprentis, souvent de fort petite taille. 27 % des établissements ont refusé, souvent à la suite d'une consultation de leur Conseil d'administration. C'était surtout de gros lycées de l'ouest parisien ou des grandes villes. Ils ont été remplacés dans leur strate.

Les jeunes enquêtés (15-18 ans) étaient tirés sur une liste anonyme d'identifiants, après stratification par âge. Les jeunes de 18 ans étaient surreprésentés par rapport à ceux âgés de 15 à 17 ans. Seuls 4 % des parents des enfants mineurs, informés par le proviseur, ont exprimé leur refus de cette enquête. Enfin 14 % des jeunes ne se sont pas présentés à l'interview organisée pendant les heures de scolarité.

Toutefois, 24 % des 15-18 ans *échappaient au champ* de l'enquête : les responsables ont jugé trop délicat d'enquêter dans les collèges les jeunes de 15 ans —qui constituent 40 % de cette génération : l'absence de cette fraction de l'univers, un

peu plus jeune ou ayant pris du retard dans leurs études, ne constitue peut-être pas un biais fondamental. Les difficultés scolaires à ces âges ne s'avèrent pas un facteur discriminant des comportements sexuels.

Comment interroger ?

Premier test ACSF : téléphone versus face-à-face

Lors des discussions préalables à l'élaboration du dispositif ACSF et de l'analyse de la littérature internationale, on évoqua la possibilité de réaliser l'enquête par téléphone. Cette suggestion rencontra des avis mitigés de la part des membres du groupe ACSF et auprès du conseil scientifique chargé de suivre la réalisation de l'enquête. On décida cependant d'explorer cette possibilité, en consultant, notamment, des équipes suisse et écossaise qui avaient déjà réalisé des enquêtes à l'aide du téléphone.

Un premier test fut réalisé en Juillet 1990 sur 800 personnes (400 en face-à-face et 400 par téléphone). Sa mise au point nécessita la réalisation de deux versions différentes du questionnaire, l'une pour un recueil classique en face-à-face, l'autre pour une utilisation à l'aide d'un système CATI (Computer Assisted Telephone Interview). Le bilan équilibré en termes de qualité des réponses entre les deux formes d'enquête permit de recourir au téléphone, méthode plus simple et moins onéreuse (ACSF Investigators, 1992). La méthode téléphonique présente en outre l'avantage d'organiser la collecte des données de manière centralisée. Elle a permis de maintenir celle-ci sous le contrôle permanent des chercheurs et des responsables des instituts de sondage qui ont ainsi pu accompagner les enquêteurs tout au long de l'enquête (Giami, et al., in Rapport ACSF 1993). Dans la mesure où les données recueillies ont été immédiatement saisies sur un fichier informatisé, cette méthode a en outre permis la collecte d'informations plus cohérentes et d'éviter les déperditions d'informations liées au transfert des données d'un fichier papier vers un fichier informatisé. Le recours au téléphone constitue une innovation dans l'étude des comportements sexuels.

Deuxième test ACSF : lettre-avis et amélioration du taux de réponse

Au moment de débuter l'enquête, l'une des grandes inconnues résidait dans son acceptabilité et donc dans le taux de réponse escompté. Il apparut rapidement que celui-ci pouvait être élevé, de l'ordre de 70% environ, si les conditions de présentation de l'enquête respectaient un certain nombre de critères : envoi d'une lettre-avis justifiant clairement l'intérêt de l'étude pour la recherche biomédicale,

identification des promoteurs et des investigateurs, perspective de Santé Publique et respect strict des réglementations en vigueur, notamment en matière de confidentialité. Un deuxième test fut ainsi réalisé en Décembre 1990 qui mit en évidence que l'envoi préalable d'une lettre-avis augmente sensiblement le taux d'acceptation de l'enquête (Riandey, Firdion, in *Population*, 1993).

Troisième test ACSF : information préalable et confidentialité

L'information préalable des répondants constitue une nécessité éthique et réglementaire. La CNIL avait envisagé de demander aux chercheurs ACSF de bien préciser dans la lettre-avis le caractère facultatif de l'enquête ainsi que de mentionner explicitement la présence de questions sur les comportements sexuels. L'inclusion de ces recommandations, dans une lettre-avis qui fut testée auprès de 300 personnes, eut pour effet d'augmenter le taux de non-réponse de 18% à 46% (Riandey, Firdion, in *Population* 1993). Dans la mesure où la lettre-avis était adressée au "ménage", il fut finalement décidé de n'informer que la personne sélectionnée (selon la méthode anniversaire) pour répondre au questionnaire, du contenu exact de l'enquête. Ce choix est justifié par la nécessité d'une part de maintenir un taux d'acceptation de l'enquête élevé et d'autre part, d'éviter que le destinataire de la lettre prenne l'initiative de refuser de répondre en lieu et place de la personne sélectionnée au sein du ménage. Ce choix permet en outre de protéger la confidentialité des réponses par rapport à l'entourage domestique du répondant.

Le taux de réponses final fut assez élevé, de l'ordre de 70%. Il fut néanmoins moins important que celui qui est habituellement observé dans d'autres enquêtes réalisées en population générale mais sur des sujets relativement moins sensibles, comme la contraception par exemple (Toulemon, Léridon, 1991).

La formation et le suivi des enquêteurs ACSF

Les enquêteurs, volontaires pour participer à cette enquête, furent formés par les membres de l'équipe ACSF (Giami et al., in Rapport ACSF, 1993) qui ont en outre assuré le suivi quotidien de la collecte sur les sites téléphoniques des instituts de sondage. Cette initiative, inscrite au cahier des charges des instituts de sondage, a permis d'établir, dès le début de la réalisation de l'enquête, des relations de collaboration avec les responsables et les superviseurs. Par ailleurs, une enquête qualitative par entretiens approfondis fut réalisée auprès des enquêteurs, avant le début de l'enquête et à la fin de celle-ci. L'analyse de ces entretiens a permis de mieux comprendre les motivations des enquêteurs, les caractéristiques de la

compétence spécifique liée à l'acte d'enquêter sur la sexualité et la dynamique de la communication sur la sexualité en situation d'enquête (Giami et al. 1997).

Choix du mode de collecte de l'enquête ACSAG

La procédure de l'entretien en face à face au domicile des personnes à interroger a été en définitive retenue aussi bien aux Antilles - où la relativement bonne couverture téléphonique laissait la possibilité de réaliser les entretiens par téléphone - qu'en Guyane, où de toutes façons l'état de la couverture téléphonique interdisait d'envisager une telle possibilité. La prise de cette décision a été grandement aidée par les résultats d'une enquête pilote, effectuée en Guadeloupe, qui comparait les deux procédures en question. En effet, d'une part, ces résultats ont montré que le taux des refus à participer à l'enquête était quatre fois, et celui des abandons en cours de passation du questionnaire trois fois, plus importants dans la collecte des données au téléphone que dans celle en face à face, et, d'autre part, l'analyse des incohérences dans les réponses recueillies dans un même questionnaire lors de cette enquête pilote a semblé indiquer une meilleure compréhension des questions et une plus grande sincérité des réponses dans la seconde procédure que dans la première. En revanche, l'analyse des pourcentages de non-réponses ou de la distribution des réponses à certaines questions-clés n'a pas donné l'avantage à l'une ou à l'autre de ces procédures quant à la fiabilité des réponses recueillies.

A la différence de nombre d'enquêtes sur les comportements sexuels réalisées dans les pays développés en face à face, qui ont prévu l'auto-administration d'une partie du questionnaire, cette possibilité a été exclue a priori en raison des multiples spécificités du rapport à la lecture et à l'écriture d'un grand nombre des habitants des départements d'Amérique : importance relative de l'analphabétisme et de l'illettrisme, diglossie, quasi-inexistence de la lecture et de l'écriture de textes en créole et dans d'autres langues vernaculaires de la région. Consciente des problèmes qu'allait engendrer ce choix, notamment la difficulté que pourraient éprouver certains enquêteurs à poser en face à face des questions de nature délicate, l'équipe ACSAG a accordé un soin particulier à la formulation des questions et à la formation des enquêteurs (voir *infra*).

Le bilan de la collecte définitive des données dans les trois départements témoigne d'une rentabilité satisfaisante des bases de sondage utilisées et, surtout, de la relative rareté des abandons et des refus à participer à l'enquête, dont les taux - même s'ils diffèrent d'un département à l'autre - sont toujours inférieurs à ceux enregistrés en Métropole (le taux de refus, par exemple, est de 13 % en Guyane, 16 % en Guadeloupe et 20 % en Martinique contre 23,5 % dans l'hexagone). Il semblerait donc que les personnes contactées ont été d'autant plus réceptives à une enquête sur les comportements que l'épidémie est sévère dans leur pays de résidence.

Pour déterminer quel individu serait interrogé au sein du ménage sélectionné, les enquêteurs d'ACSAG, comme ceux d'ACSF, ont utilisé la méthode de l'anniversaire, jugée plus facile à mettre en oeuvre que les autres méthodes possibles. Cependant, l'emploi de cette méthode s'est révélée problématique dans une minorité de cas.

En effet, d'une part, la taille des ménages est plus importante dans les départements d'Amérique que dans l'hexagone et la personne qui était contactée par l'enquêteur n'avait pas toujours en mémoire les dates d'anniversaire de tous les membres de leur ménage. D'autre part, un certain nombre de personnes contactées déclaraient ne pas fêter les anniversaires et donc ignorer les dates de naissance des autres membres du ménage. En général, ce problème était résolu par l'enquêteur en consulant le livret de famille ou parce que la personne contactée arrivait en définitive à situer au moins le mois de naissance de l'ensemble des adultes qui habitent avec elle.

La formulation des questions : linguistique et sociologie du langage

La formulation des questions à thème sexuel constitue un problème important. Faut-il choisir les termes dans le registre médical, dans un langage de sens commun ou bien avoir recours à des termes pouvant être considérés comme obscènes et risquant ainsi de heurter la sensibilité des répondants (et des enquêteurs !). Il fut demandé à une équipe de linguistes et de sociologues du langage de réfléchir à ce problème. Quelques questionnaires de l'enquête pilote ACSF par téléphone furent ainsi enregistrés. Les linguistes mirent ainsi en évidence que la formulation des termes sexuels ne constitue pas le problème majeur à résoudre. En effet, il n'existe pas de formulation neutre en la matière. L'analyse des questionnaires enregistrés contribua à la compréhension de la dynamique langagière de la passation du questionnaire. Celle-ci doit être comprise comme une forme de conversation asymétrique dans laquelle le sondé parle beaucoup moins que l'enquêteur. La passation du questionnaire apparaît donc ainsi comme une situation paradoxale dans laquelle "le sondé ne se contente pas de répondre aux questions mais se trouve engagé avec l'enquêteur dans la tâche commune de remplir le questionnaire » (Achard, 1994). La complexité du questionnaire et sa longueur auraient donc favorisé l'instauration d' une relation de partenariat avec les répondants de l'enquête.

Le problème du langage s'est posé de manière encore plus complexe à l'équipe ACSAG, compte tenu des fortes particularités langagières et linguistiques des pays antillais et guyanais. Il a ainsi fallu que cette équipe substitue fréquemment aux termes "techniques" du questionnaire métropolitain des expressions plus conformes au niveau de langue du français communément employé dans ces pays ou, au moins, propose aux enquêteurs des reformulations possibles de ces termes. Mais, surtout, le bilinguisme créole/français a rendu nécessaire de décider que l'enquête pourrait être

également passée en créole, ne serait-ce que pour pouvoir interroger les créolophones unilingues que le sort désignerait pour être enquêtés. En conséquence, une version dans cette langue du questionnaire a été préparée, puis son utilisation expérimentée lors des enquêtes pilotes. Ces dernières ont révélé que réaliser les entretiens en créole posait quelques difficultés aux enquêteurs, essentiellement à cause de la réticence de certaines personnes à être ainsi interrogées. C'est que, malgré la dominance de la langue créole dans les échanges verbaux de la vie quotidienne des Antilles et de la Guyane, cette langue reste socialement minorée. Dès lors, mener une enquête scientifique sur un sujet aussi sensible que celui de la sexualité dans une langue qui est celle de l'intimité affective et de la proximité sociale ou encore celle des ordres donnés par un maître à ses subordonnés peut susciter des réactions négatives importantes chez ceux à qui on s'adresse ainsi. Comme, par ailleurs, les enquêtes pilotes ont établi que, dans la très grande majorité des cas, la maîtrise que les personnes interrogées ont du français leur permettait de comprendre sans difficulté la plupart des questions qui leur étaient posées et d'y répondre clairement, il a été décidé que, sauf contre-indication manifeste, la passation du questionnaire serait au départ proposé en français aux enquêtés et que le passage au créole se ferait, si besoin en était, en recourant à la version préparée à cet effet. Enfin, la formation des enquêteurs, dont la quasi totalité étaient des personnes parlant le créole, a dispensé, sous forme d'exercices et de jeux de rôles, une préparation spécifique à la passation du questionnaire dans cette langue.

L'effet enquêteur

Les résultats de l'enquête qualitative menée sur les enquêteurs ACSF ont mis en évidence que le maniement d'un questionnaire de cette nature et de cette longueur repose sur de fortes motivations d'une part, et d'autre part, suscite un retentissement fantasmatique important chez ceux-ci. Le suivi quotidien des enquêteurs a contribué à canaliser leurs réactions envers l'enquête et envers les répondants. Dans une telle situation, où la subjectivité des enquêteurs a été fortement mobilisée, on n'aurait pas été surpris d'observer des biais liés aux caractéristiques des enquêteurs. L'analyse qui a été réalisée sur les résultats (Firdion et Laurent, rapport scientifique, 1997) n'a fait apparaître que quelques biais limités liés au sexe de l'enquêteur à propos des opinions concernant certaines pratiques sexuelles. Le questionnaire et sa passation par téléphone auraient-ils eu pour effet de réduire la complexité de la communication sur la sexualité liée à la différence des sexes ? L'étude de l'effet enquêteur illustre bien le type de collaboration qui a pu s'établir entre chercheurs travaillant à l'aide de méthodes quantitatives et qualitatives.

Quelles Questions Poser

La structure du questionnaire ACSF

Le schéma de base a été élaboré à partir d'une revue de la littérature existante et orienté par la problématique du risque. Le questionnaire a été construit autour d'un noyau dur comprenant une entrée à partir des caractéristiques du répondant, sa biographie sexuelle, une description détaillée du "dernier rapport sexuel" quel que soit le partenaire avec lequel il s'est déroulé et éventuellement de l'avant-dernier rapport si le répondant avait eu plus d'un partenaire au cours des douze mois précédant l'enquête. Le questionnaire ACSF comporte en outre un certain nombre de "modules" qui reflètent des problématiques spécifiques. Ces modules sont présentés ici dans l'ordre chronologique de leur apparition dans le déroulement du questionnaire.

STRUCTURE DU QUESTIONNAIRE ACSF
Signalétique
Divers santé
CARTE FILTRE
Parler de sexualité
Normes concernant la sexualité
Normes de l'acte sexuel
Attitudes et Perceptions de la mort
Attitudes temporelles
Prise de risque
"Locus of control"
Normes concernant l'usage de préservatifs
Confidents
Vie de couple
Contraception
Caractéristiques: religion, revenu , lecture
Premier rapport
Pratiques sexuelles générales
Nombre de partenaires
Dernier rapport sexuel
Avant-dernier rapport sexuel
Expérience de la prostitution
Maladies Sexuellement Transmissibles
Test de dépistage
Changements de comportements (passé-futur)
Perception du risque
Usage de drogues illicites
Abus sexuels
Connaissance et facilité d'utilisation des moyens de protection
Perception sociale
Fantasmes
Connaissance séropositif
Solidarité

Le questionnaire n'a donc pas été élaboré selon une problématique unifiée. Son élaboration répond à un certain nombre de préoccupations théoriques qui ont été appliquées à l'étude de l'activité sexuelle en réponse à la commande de Santé publique :

- Contextualisation de l'activité sexuelle : L'étude de la relation entre les partenaires, (Leridon, in *Population*, 1993, Messiah, Pelletier, 1996) met en évidence que l'activité sexuelle est "négociée" - explicitement ou implicitement - selon le "type de partenaire" avec lequel on se trouve.

- Réseaux sociaux des individus : l'ACSF constitue, en France, la seule enquête téléphonique sur un large échantillon à avoir abordé le thème des réseaux personnels, et notamment des réseaux de confidence (Ferrand, Mounier, in *Population* , 1993 ; Ferrand, in Rapport scientifique à paraître). Cette approche se situe dans une problématique plus vaste de la communication autour du thème de la sexualité. Les transactions entre les partenaires, ainsi que la communication dans l'enfance ont été abordées.

- Normes et représentations sociales de la sexualité : l'activité sexuelle est une activité qui est sujette à des normes sociales très fortement intériorisées. La perspective d'étude du changement de certaines pratiques sexuelles impliquait donc l'étude des normes sociales de la sexualité (Spencer, in *Population,* 1993).

- Construction du temps et dynamique du changement : l'activité sexuelle est conçue comme évolutive au cours des cycles de vie déterminés par des facteurs biologiques, psychologiques et relationnels selon une temporalité qui a permis de distinguer : le cours de la vie, les 5 dernières années, les 12 derniers mois, les 4 dernières semaines, "actuellement". Par ailleurs, la demande de santé publique impliquait l'évaluation de changements réalisés ou en cours de réalisation "depuis l'apparition du sida". On a ainsi pu évaluer les changements subjectifs, c'est à dire considérés comme tels par les individus et les changements de pratique explicables par la place dans le cycle de vie ou par une décision aux ressorts certes complexes.

- Homosexualité, bissexualité et hétérosexualité : demander de préciser lors de nombreuses questions le sexe du partenaire a remis en cause le caractère "évident" de l'hétérosexualité et le clivage entre homosexualité et hétérosexualité. Le questionnaire a ainsi permis de mettre en évidence différentes figures de l'hétérosexualité, de l'homosexualité et de la "bisexualité". (Messiah, Mouret-Fourme, in *Population,* 1993, et 1995).

- Utilisation du préservatif : la mise en relation de l'utilisation du préservatif avec le type de partenaire sexuel et l'étude des motivations liées à son usage (contraception ou prévention) ont permis de renouveler la problématique des enquêtes de contraception et de mettre en évidence les significations multiples du préservatif

comme moyen contraceptif ou comme moyen de prévention selon le contexte et le type de partenaire avec lequel il est utilisé.

- Fantasmes sexuels : L'étude des fantasmes qui est classique pour les psychanalystes (Stoller, 1984), est plus rarement abordée par les sociologues et encore moins dans le champ de la santé publique (Lisandre, 1996). Le module consacré aux fantasmes sexuels, dans le questionnaire, illustre la migration des thématiques de recherche d'un champ à un autre en permettant la formulation de problématiques de recherches originales.

L'analyse comparative et historique du questionnaire a mis en évidence l'influence d'une "représentation épidémiologiste de la sexualité" liée, à la problématique du sida (Giami, 1991, in *Population,* 1993). Cette influence est repérable notamment par la présence de nombreuses questions concernant l'homosexualité masculine, les pratiques anales, l'évolution des nombres de partenaires, l'usage de drogues, l'utilisation du préservatif. Par ailleurs, les pratiques sexuelles sont explorées principalement comme des formes de contact entre muqueuses. Inversement, le questionnaire comprend peu de questions sur la masturbation (Béjin, in *Population,* 1993) et ne comprend pas de questions sur l'avortement ni sur les positions sexuelles. Les violences sexuelles ont en outre fait l'objet d'une exploration (Bozon, in Rapport ACSF, 1993).

La structure du questionnaire de l'enquête ACSAG est sensiblement la même que celle de celui d'ACSF.

Le questionnaire ACSJ

L'entretien de l'enquête ACSJ durait en moyenne une heure et quart. Il ne comportait pas de section auto-administrée pour les questions les plus personnelles. Ce questionnaire était en effet fondé sur une suite de filtres permettant d'éviter au jeune d'être interrogé sur des types de pratiques qu'il n'a pas eues. Notons qu'un questionnaire auto-administré ne facilite pas le respect des filtres et ne garantit pas le secret sur les questions sensibles non posées. Au contraire, le système de filtres prévoyait des questions de contrôle : ainsi le questionnaire récupérait les jeunes qui déclaraient ne pas avoir eu de rapport sexuel, mais déclaraient des caresses génitales.

Le questionnaire reposait aussi sur un enchaînement de concepts à manier avec précision. L'entrée dans la vie sexuelle est analysée comme un processus en trois étapes : premier baiser, premières caresses, premier rapport pénétratif. L'étalement de ces étapes et le changement de premier partenaire d'une étape à l'autre sont d'une grande signification.

Le pari sur la sincérité des réponses s'appuyait sur la dimension biographique du questionnaire : il est difficile de rester cohérent en s'inventant une biographie dont les éléments ne sont annoncés que progressivement.

Outre cette dimension biographique, le questionnaire décrivait les réseaux affectif ou sexuel à travers les pairs, les copains et les partenaires. Le questionnaire repérait les individus cités par un numéro permettant de les retrouver dans les différentes sections du questionnaire. Ce système efficace comporte des risques d'erreur —bien connus dans les enquêtes pointues— et a contribué aux 4 à 5 mois d'apurement du fichier. Ainsi les erreurs d'appariement avaient conduit à 400 partenaires du même sexe que l'enquêté. Après corrections des appariements, le taux de pratique homosexuelle n'était plus voisin de 6 % mais s'établissait à 1,3 %. Cet exemple illustre bien la nécessaire *exigence critique à l'égard des données* que l'équipe ACSJ s'est imposée.

Analyser les données

Le recours aux méthodes classiques de la statistique a été très important, mais ce fut également l'occasion d'appliquer à des données de comportement et de santé, des méthodes d'analyse qui jusque là avaient été réservées à d'autres domaines. Il en est ainsi par exemple de l'analyse de l'évolution temporelle du nombre de partenaires sexuels présentée dans le prochain rapport scientifique par Henri Leridon. La méthode de constitution d'un indicateur synthétique de signification longitudinale à partir de données recueillies de façon transversale est classique dans le champ de la démographie. Sa transposition simple et originale aux données concernant l'activité sexuelle permet de suspecter que des modifications importantes pourraient être en train de s'opérer sous nos yeux (en l'occurrence une diminution du nombre de partenaires sexuels avant la constitution des couples stables) et que seules des études longitudinales ou transversales répétées dans le temps permettront de vérifier.

Il en va de même de la réalisation, par Françoise Lepont d'un modèle de simulation événementiel de la dynamique de l'épidémie. A partir des estimations fournies de façon relativement précise par l'enquête, il est en effet possible de simuler l'évolution des comportements d'une fraction de la population et d'en évaluer les conséquences sur la dynamique de l'épidémie. A partir de ce modèle, on peut tenter d'anticiper l'effet attendu de modifications qui pourraient être induites par des stratégies de prévention et ainsi contribuer aux choix importants dans le domaine de la Santé Publique. Il faut cependant prendre en compte les limites de l'échantillon réuni par l'ACSF : celui-ci n'a pas permis d'inclure un nombre suffisant d'usagers de drogues, ni de personnes séropositives pouvant être insérées dans la construction du modèle. L'élaboration de modèles prospectifs de l'épidémie ne peut reposer que sur l'analyse des comportements, sans prendre appui sur les données socio-

épidémiologiques qui indiquent la répartition sociologique de l'épidémie de VIH. L'application d'un tel modèle doit donc être interprétée avec prudence.

Les résultats de l'enquête ACSAG en ce qui concerne l'intensité (en termes de pourcentage des sexuellement actifs ou de fréquence des rapports) ou la précocité de l'activité sexuelle ou encore les pratiques sexuelles (à l'exception de la masturbation) sont des plus semblables à ceux de l'enquête ACSF.

Les seules spécificités notables des caractéristiques de l'activité sexuelle dans les départements d'Amérique dont témoigne l'enquête ACSAG sont l'importance de la fréquentation masculine de prostituées en Guyane et, surtout, la fréquence mais aussi la nature du multipartenariat hétérosexuel dans ces départements. Globalement, le pourcentage d'hommes vivant aux Antilles ou en Guyane qui ont déclaré avoir eu plus d'un partenaire durant l'année précédant l'enquête est en Martinique plus du double, le triple ou plus du tripe en Guadeloupe ou en Guyane, de ce qu'il est dans l'hexagone. En revanche, les femmes de ces départements ne sont - au moins aux Antilles - que légèrement plus nombreuses à déclarer être multipartenaires que les femmes de la France métropolitaine. Ce multipartenariat n'est pas, aux Antilles et en Guyane, l'apanage des jeunes mais davantage un comportement qui persiste durablement chez les individus, même si son importance tend - comme partout ailleurs - à s'amenuiser avec le vieillissement. Par ailleurs, la part du multipartenariat simultané (mesuré sur le nombre de partenaires actuels) est particulièrement forte chez les hommes des DFA, y compris chez ceux vivant en couple. Enfin, le multipartenariat englobe aux Antilles une forte proportion d'individu dont toutes les partenaires sont des personnes avec lesquelles ils faisaient déjà l'amour avant l'année précédant l'enquête (multipartenaires stables) : ils représentent une moitié environ des hommes multipartenaires des Antilles contre 17 % en Guyane et 30 % en Métropole.

L'essentiel est qu'il existe un parallélisme étroit entre les données comportementales de l'exposition au risque d'infection par le VIH et les données épidémiologiques de la prévalence de cette infection.

Contexte de l'enquête et fiabilité des données

Le dispositif de l'enquête ACSF a permis d'examiner la qualité des réponses en fonction du contexte de l'entretien. Le dispositif de l'enquête prévoyait de rappeler automatiquent les numéros téléphoniques jusqu'à douze fois, en cas de non réponse à l'appel. Or les personnes qui n'ont pu être jointes qu'au delà du sixième appel téléphonique se sont avérées être significativement plus souvent des multipartenaires que celles qui ont eu un "rang d'appel" moins élevé. Cette "obstination à rechercher les répondants" s'est ainsi montrée utile. Inversement, la présence du conjoint lors de la passation du questionnaire, semble avoir suscité une sous-déclaration du

multipartenariat et de la consommation de drogues. En outre, les enquêteurs ont alors eu plus souvent l'impression que les réponses étaient "moins sincères". (Firdion, in *Population*, 1993).

Dans l'enquête ACSAG, une part encore plus infime des enquêtés ont déclaré avoir eu, au cours de leur existence, des partenaires du même sexe qu'eux (un peu moins de 1 % des hommes et de 2 % des femmes).

La faiblesse relative du nombre d'homosexuels et de bisexuels que comportent nos échantillons antillais et guyanais est-elle le fidèle reflet de la réalité ? L'homosexualité et la bisexualité sont-elles au contraire massivement tues dans ces pays ? Le fait que ces comportements y sont très fortement condamnés, surtout en ce qui concerne l'homosexualité masculine, pourraient le laisser penser : ainsi, aux Antilles, autour de 10 % seulement des personnes des deux sexes jugent acceptable que deux hommes fassent l'amour ensemble, contre 30 % des hommes et un peu plus de 40 % des femmes en Métropole. Mais comment comprendre alors qu'un rejet de force inégale, selon le département et le sexe des enquêtés puisse avoir provoqué une "sous-déclaration" de niveau identique dans les trois départements d'enquête et pour les deux sexes ?

Différences de pratique ou de déclaration

Pour de nombreuses pratiques sexuelles, la fréquence diffère peu d'un milieu social à l'autre (Bozon, in Rapport ACSF, 1993). Ainsi parmi les hommes de moins de 50 ans, dans tous les groupes sociaux, les trois quarts des individus déclarent avoir pratiqué le cunnilingus. Cette tendance à l'uniformisation dans les pratiques est cependant contrecarrée par certains processus sociaux et culturels. Par exemple, chez les femmes des générations correspondantes, la fellation et la masturbation sont plus souvent mentionnées dans les groupes aisés, ce qui indique sans doute à la fois une difficulté de déclaration et une moindre pratique dans les milieux populaires. Inversement, on peut citer l'exemple de la pénétration anale, un peu plus fréquemment déclarée dans certains milieux populaires, ouvriers non qualifiés, et femmes employées de commerce et de service ; l'enquête ACSJ indique d'ailleurs également une expérience plus grande de cette pratique parmi les apprentis et les lycéens d'enseignement professionnel. Les comportements sexuels comportent ainsi une dimension culturelle, comme le confirme, par exemple, l'écart considérable (de 20 à 40 points) entre les pourcentages enregistrés dans les DFA et en Métropole en ce qui concerne la masturbation. Cet écart témoigne d'une très forte dévalorisation de l'auto-érotisme dans ces départements.

On constate de nombreuses différences entre les comportements déclarés des hommes et ceux des femmes, que l'on n'arrive pas toujours à interpréter de façon satisfaisante : hommes et femmes déclarent des nombres moyens (et médians) de

partenaires sexuels au cours de leur vie qui ne coïncident pas. Ainsi, en Métropole, les hommes déclarent en moyenne avoir eu 12 partenaires (avec une médiane de 5). Les femmes en déclarent 3 en moyenne (avec une médiane de 1). Même la prise en compte de la prostitution féminine ne suffit pas à expliquer cet écart.

Selon une hypothèse plausible, les enquêtés des deux sexes auraient envisagé différemment ceux qu'elles comptaient comme des partenaires : les femmes auraient probablement compté les hommes avec lesquels elles avaient noué des relations d'une certaine durée, tandis que les hommes auraient dénombré toutes celles avec lesquelles ils avaient eu des expériences sexuelles, aussi brèves soient elles. Apparemment, les femmes préfèrent oublier les expériences trop peu marquantes, ne correspondant pas à leur idée de ce qu'une relation doit être, intériorisant ainsi des attentes sociales très contraignantes à leur égard.

On enregistre des résultats beaucoup plus proches entre hommes et femmes si l'on compare les nombres de partenaires déclarés dans les cinq dernières années, ou dans les douze mois. Ces dernières déclarations relèveraient du comptage, alors que le nombre de partenaires sur la vie renverrait plus à la représentation d'une identité.

On observe ces mêmes divergences dans à l'enquête ACSAG, notamment celles concernant les divergences qui se manifestent entre les hommes et les femmes sur le nombre de leurs partenaires sexuels au cours de la vie, et le recours à l'auto-masturbation des femmes. Ces divergences sont d'autant plus importantes à comprendre que nombre de résultats des analyses des données collectées témoignent d'une bonne fiabilité des réponses à l'enquête.

Ainsi, le fait que, à l'exception des questions portant sur les fantasmes, le taux de non-réponse par question reste si faible tout au long du questionnaire permet d'écarter l'éventualité que nos interrogations aient fréquemment et fortement suscité une incompréhension ou une résistance.

La relativement bonne cohérence des réponses d'un même enquêté tout au long du questionnaire, tout comme la fréquente concordance des réponses entre hommes et femmes vont aussi dans le sens de cette bonne fiabilité.

Éthique

La réalisation d'une telle enquête pose-t-elle les questions éthiques communes à l'ensemble des recherches scientifiques ou bien pose-t-elle des questions éthiques spécifiques et liées à son objet, la sexualité ?

Ainsi toutes les exigences de la loi Huriet-Sérusclat du 20 Décembre 1988 ont été respectées. L'information sur les objectifs de la recherche a été progressivement

divulguée aux répondants potentiels et aux personnes retenues pour répondre à l'enquête. Le consentement préservant l'autonomie du répondant - sa capacité - à interrompre à tout moment le déroulement du questionnaire a été recueilli. La confidentialité des données a été préservée grâce à l'anonymat de la procédure de recueil. Dès que le répondant formulait sa première réponse au questionnaire, il devenait impossible, même en cas de rupture intempestive de la communication téléphonique d'identifier celui-ci et en particulier de le rappeler pour terminer l'enquête. Les règles qui doivent encadrer toute recherche scientifique menée auprès de sujets humains ont donc été respectées lors de ce travail. La réalisation de l'enquête ACSF a démontré qu'une telle investigation est possible et qu'elle est acceptable par une grande partie de la population générale.

La question éthique ne saurait cependant se résoudre à l'application des réglementations existantes :

L'acceptation du principe de l'enquête par les répondants soulève une première question. Doit-on considérer que répondre à un questionnaire détaillé sur la sexualité relève d'une évolution de la liberté de parole, ou bien qu'elle constitue une forme de soumission à l'autorité médicale qui a donné sa légitimité à la réalisation d'une telle enquête ?

La connaissance de la sexualité humaine reste entourée d'un tabou et suscite des prises de position politiques. La recherche scientifique dans ce domaine n'a pu se développer, la plupart du temps, que lorsque surgit un "problème" qui vient légitimer que l'on puisse s'intéresser à la sexualité. La responsabilité des chercheurs se trouve donc engagée dans la mesure où les résultats de leur activité peut contribuer à une réduction des souffrances humaines.

De plus, l'existence de stéréotypes sur la sexualité de différents groupes peut rendre problématique la communication des résultats d'enquêtes sur les comportements sexuels, lorsque ceux-ci paraissent donner du crédit à de telles idées reçues. Par exemple, la confirmation de l'importance du multipartenariat hétérosexuel dans les DFA et celle du recours des hommes aux services de prostituées en Guyane ont suscité un malaise dans certains secteurs de l'opinion publique antillaise et surtout guyanaise : les données de l'enquête ACSAG ont été suspectées de porter la marque des stéréotypes racistes sur une prétendue "sexualité nègre" (et ce d'autant que ces données étaient présentées en contraste avec celles concernant la Métropole). Une telle situation a le mérite de nous rappeler aux chercheurs leurs responsabilités quant aux effets pervers éventuellement suscités par la communication de leurs résultats. Ainsi, l'équipe ACSAG n'a eu de cesse de rappeler que le multipartenariat est un phénomène universel, même si son ampleur et ses caractéristiques diffèrent souvent d'un groupe à un autre en fonction des histoires particulières et des cultures singulières de ces groupes.

Des questions portant sur des pratiques sexuelles minoritaires ou considérées comme "déviantes" ont pu "choquer" certains répondants. Le principe d'égalité, ainsi que les objectifs scientifiques de la recherche, a conduit les chercheurs de l'ACSF à poser les mêmes questions à l'ensemble des personnes retenues pour le questionnaire long,qu'elles soient supposées à risque ou non à risque. Le seul filtre retenu concernant les questions sur les pratiques sexuelles a porté sur l'expérience de rapports sexuels au cours de la vie ou sur l'orientation sexuelle qui implique des pratiques spécifiques.

Au contraire, pour l'enquête ACSJ réalisée auprès de mineurs, il a été jugé essentiel que les questions posées ne révèlent pas des pratiques inconnues des jeunes peu encore engagés dans la vie sexuelle.

BIBLIOGRAPHIE

Une bibliographie détaillée figurant dans le prochain rapport scientifique peut être demandée aux auteurs.

Spira, A., Bajos, N., Groupe ACSF, *Les comportements sexuels en France,* Rapport ACSF 1993, Paris, La documentation française.

S. d. M. Bozon et H. Leridon (1993), "Sexualité et Sciences sociales. Les apports d'une enquête", *Population*, n° 5.

Acsf Bajos, N., Bozon, M., Giami, A., Doré, V., & Souteyrand, Y. (Eds). (1995), *Sexualité et sida : recherches en sciences sociales*, Paris, ANRS.

ACSF investigators (1992), "AIDS and sexual behaviour in France", *Nature*, 360(3):407-409.

Bozon, M., Léridon, H., & Riandey, B. (1993), "Les comportements sexuels en France : d'un Rapport à l'autre", *Population et Sociétés* , 276, 1-4.

Giami, A. (1991), "De Kinsey au sida: l'évolution de la construction du comportement sexuel dans les enquêtes quantitatives", *Sciences Sociales et Santé*, vol IX, 4, 23-56.

Giami, A., (1993), "Le questionnaire de l'enquête ACSF - Influence d'une représentation épidémiologique de la sexualité", *Population*, 5, 1229-1256.

Giami, A., (1996), "Non response and Don't know answers in sex surveys", *Social Science Information* , 35, (1), 93-109.

Giami, A., Schiltz, M.-A. (1996), "Sex Research in France in the era of AIDS. Representations of sexuality and relations between partners", *Annual Review of Sex Research,* à paraître.

Kinsey, A., Pomeroy, W., & Martin, C. (1948), *Sexual Behavior in the Human Male,* Philadelphia, Saunders.

Laumann, E., Gagnon, J.H., Michael, R., & Michaels, S. (1994), *The Social organization of Sexuality. Sexual Practices in the United States*, Chicago, The University of Chicago Press.

Moatti J.P., Dab W., Pollak M. & Groupe KABP France (1992), "Les Français et le SIDA", *La Recherche*, 247, 1202-1211.

Moatti, J.-P., Grémy, I., Obadia, Y., Bajos, N., Doré, V. & Groupe KABP/ACSF France (1995). "Sida: Dernière enquête nationale. Succès et risques des campagnes de prévention", *La Recherche*, 282, 30-34.

Simon, P., Gondonneau, J., Mironer, L., & Dourlen-Rollier, A.-M. (1972), *Rapport sur le comportement sexuel des Français,* Paris, Julliard, Charron.

Enquête ACSAG :

M. Giraud, A. Gilloire, P. de Colomby et S. Halfen, *Analyse des comportements sexuels aux Antilles et en Guyane*, ANRS, juillet 1994, document multicopié.

M. Giraud, A. Gilloire, S. Halfen et P. de Colomby, "Analyse des comportements sexuels aux Antilles et en Guyane" in N. Bajos, M. Bozon et A. Giami, *Sexualité et Sida. Recherches en sciences sociales*, Paris, ANRS, 1995, p. 271-275.

M. Giraud, A. Gilloire, S. Halfen et P. de Colomby, "Sexualité et Sida Aux Antilles et en Guyane", *ANRS information*, 17, janvier 1996, p. 28-36.

M. Giraud, A. Gilloire, S. Halfen et P. de Colomby, "Los comportamientos sexuales en el Caribe frances", *Estudios demograficos y urbanos,* Colegio de Mexico, sous presse.

M. Giraud, "Comportamientos sexuales y prevencion del sida en el Caribe frances", *Estudios demograficos y urbanos,* Colegio de Mexico, sous presse. (ACSAG)

Enquête ACSJ

Lagrange, H., Lhomond, B., Calvez, M., Darsch, C., Favier, C., Fierrp, F., Levinson,S., Maillochon, F., Mogoutov, A., Roché, S., & Warszawski, J. (1996), "Enquête sur les comportements sexuels des jeunes de 15 à 18 ans", *ANRS Information*, 17, 14-27.

OBJECTIFS ET MÉTHODOLOGIE DES ENQUÊTES AUPRÈS DES SANS-DOMICILE À PARIS

Maryse Marpsat et Jean-Marie Firdion (Ined),
Viviane Kovess et Caroline Mangin-Lazarus
(Association l'Elan retrouvé)

1. Le contexte des enquêtes

L'Institut National d'Etudes Démographiques (Ined) et l'Association l'Elan retrouvé ont réalisé pendant les hivers 1994-1995 et 1995-1996 deux enquêtes[1] auprès d'un échantillon représentatif de personnes sans domicile utilisatrices de services d'hébergement et de distribution de nourriture, à Paris. Nous en exposons la méthodologie largement commune, inspirée des méthodes de sondage et d'estimation d'enquêtes américaines (Urban Institute, Research Triangle Institute). Il se pose bien sûr un problème complexe d'échantillonnage puisque la population échappe à la base de sondage classique des logements, mais de telles enquêtes révèlent des questions de fond comme d'éthique, avant même de rencontrer les difficultés pratiques qu'on imagine.

1a. Comment se présente la question des sans-domicile en France ?

La France des années cinquante avait déjà connu une importante crise du logement. La crise actuelle semble de nature différente. Elle se situe dans un contexte de crise économique et de fragilisation des emplois et dans une période où les liens familiaux et sociaux sont aussi plus fragiles ; ces facteurs se combinent avec une disparition progressive du "parc social de fait", c'est-à-dire des logements inconfortables mais bon marché, qui sont détruits ou rénovés avec un remplacement de leurs habitants par des personnes plus aisées.

1 Les travaux de l'Ined ont été réalisés sous l'égide du CNIS, en partenariat avec les associations et les autres organismes se préoccupant des sans-domicile, et avec le concours de la Commission des Communautés Européennes, du Ministère de l'Equipement, du Ministère des Affaires sociales, de la Santé et de la ville, et de la Fondation Abbé Pierre.

Différents observateurs s'accordent à relever la montée du nombre des sans-abri en France, ainsi que les modifications de certaines de leurs caractéristiques : ce seraient plus qu'auparavant des femmes, des jeunes, des immigrés en provenance de pays dont ils fuiraient la situation politique ou économique (pays de l'Est en particulier ces dernières années). Toutefois, cette population fortement hétérogène demeure mal connue en raison des difficultés méthodologiques auxquelles se heurtent les approches statistiques. En particulier, si les différentes estimations "sauvages" fleurissent, il n'existe pas actuellement d'estimation fiable de leur nombre au niveau national.

C'est pourquoi, en 1993, la réalisation d'une enquête nationale fut demandée, au sein du Conseil National de l'Information Statistique (CNIS), par les associations et certains syndicats de salariés. Il s'agissait de disposer d'informations qualitatives et quantitatives, pour préciser l'ampleur du phénomène et repérer non seulement les personnes exclues du logement mais aussi celles menacées de l'être.

1b. Objectifs des enquêtes

L'enquête de l'Ined auprès de la clientèle des services destinés aux sans domicile réalisée à Paris s'insère dans ce projet plus vaste. Dans le but de proposer "un schéma d'investigation qui permette de connaître de façon aussi scientifique que possible la situation des personnes exclues du logement, les processus qui les ont conduites à cette exclusion et les obstacles qu'elles rencontrent dans la recherche d'un logement", plusieurs enquêtes expérimentales se sont déroulées en France pendant l'hiver 1994-1995, dans des zones géographiques réduites :

- une enquête sur les ménages menacés d'expulsion

- des enquêtes sur les sans domicile "au sens strict" (dont celle de l'Ined)

- des enquêtes sur les conditions de logement des ménages à très bas revenus.

L'expérimentation de l'Ined avait comme objectif essentiel d'étudier les conditions de réalisation et d'extension à une plus grande échelle d'enquêtes fondées sur les méthodologies américaines et d'élaborer des recommandations en vue d'aboutir à une meilleure prise en compte des personnes sans domicile dans le recensement et les enquêtes classiques. Il en était attendu également une première exploration des processus conduisant à la situation de sans-domicile.

L'enquête de l'Elan retrouvé a été effectuée durant l'hiver 1996, de février à mi-mars, auprès d'un échantillon représentatif constitué selon la méthode de l'Ined. Elle avait pour objectif de connaître la prévalence des principaux problèmes de santé mentale dans la population des sans-abri de Paris. Elle doit aussi permettre de

décrire cette population suivant quelques axes socio-démographiques, l'histoire de son itinérance, la présence de maladies somatiques sévères et de handicaps et l'utilisation des soins. A cette fin ont été adaptés les principes de la méthodologie clinique de l'enquête effectuée sur le même sujet par Louise Fournier à Montréal (1990).

Durant l'hiver 1995, une enquête pilote réalisée sur un échantillon de 100 personnes avait permis de tester le contenu d'un questionnaire sur la santé mentale et la pertinence des résultats grâce à des entretiens menés en double par un psychiatre pour chaque questionnaire.

A terme la connaissance de la prévalence des problèmes de santé mentale et de leur prise en charge doit permettre d'éclairer les prises de décision quant à l'organisation de l'accès aux soins de cette population spécifique.

2. La méthode employée

2 a. Les études réalisées aux États-Unis

On doit distinguer plusieurs générations d'études sur les personnes sans domicile développées aux Etats-Unis, depuis le début des années quatre-vingt. (Burt, 1992 ; Firdion et Marpsat, 1994) :

Les premières, vivement contestées, sont fondées sur des opinions d'experts : il en est ainsi des chiffres avancés par la Community for Creative Non Violence (CCNV) (Hombs et Snyder, 1983), qui lancèrent le débat national sur le nombre de sans-abri, et de l'estimation du HUD (US Department of Housing and Urban Development, 1984). Etablie avec des méthodes plus explicites, quoique reposant également en partie sur des opinions d'experts, cette dernière estimation fut vivement contestée par les militants en faveur des sans-abri.

Les enquêtes de la deuxième génération se déroulaient de nuit, simultanément dans la rue et les centres d'hébergement. Cette méthode de collecte ,mieux fondée, a été tentée par le Bureau du Census américain en 1990, mais présente de grosses difficultés pour son volet "rue".

La troisième génération d'enquêtes se déroule de jour et sur une période plus longue, dans les "services" destinés aux personnes en difficulté. Citons les travaux de, Martha Burt de l'Urban Institute sur un échantillon national de villes et ceux de Michael Dennis au Research Triangle Institute, dont l'enquête était le volet « sans domicile » d'une enquête sur la consommation de stupéfiants portant sur l'ensemble de la population de l'aire métropolitaine de Washington. L'élimination des doubles comptes et le calcul de pondérations constitue la principale difficulté de ce type

d'enquête. Il s'agit en effet de corriger les différences dans les probabilités individuelles d'insertion dans l'échantillon, dues à une utilisation plus ou moins intense des services. C'est cette méthode que reprend le Bureau du Census pour l'enquête nationale auprès des sans domicile de novembre 1996.

2 b. L'échantillonnage

Le champ des enquêtes

L'enquête de l'Ined s'adresse aux personnes majeures sans domicile « au sens strict », c'est-à-dire dormant dans des centres d'hébergement (urgence ou longue durée[2]) ou dans « la rue » (y compris parkings, gares, et autres lieux non prévus pour l'habitation). Elle est représentative de celles de ces personnes qui utilisent à Paris les services d'hébergement et de distribution de nourriture (y compris soupe et café, la nuit) à destination des personnes sans domicile. Toutefois, nous avons voulu enquêter aussi auprès des personnes dans des situations proches rencontrées dans les centres de distribution de nourriture : les personnes en squat ainsi que celles qui se déclarent hébergées mais pas de façon régulière, répondent au même questionnaire que les personnes n'ayant pas de logement. L'entretien débute donc par un questionnaire filtre qui n'élimine personne de l'enquête, mais sert à orienter sur des questionnaires appropriés à la situation par rapport au logement. En réalité, il n'y a pas une frontière nette entre avoir ou non un logement, mais plutôt un continuum de situations.

Cependant l'enquête ne couvre pas les personnes dormant dans la rue ou les squats, ou hébergées chez des amis, qui n'utiliseraient pas les centres de distribution de nourriture. Un test réalisé une nuit auprès de personnes dormant dans la rue, nous permet de penser que le nombre de ces dernières qui n'utilisent jamais aucune distribution de repas est faible, sauf peut-être parmi les plus jeunes. Ce résultat est conforté par une série d'entretiens approfondis, menés sur une période de trois mois, auprès de personnes dormant à la rue[3] et par l'enquête de l'Elan retrouvé.

L'enquête a été réalisée auprès des utilisateurs des centres de Paris intra-muros, auxquels s'ajoutent le Centre d'accueil et de soins hospitaliers de Nanterre et le centre Corentin Celton d'Issy-les-Moulineaux, où sont conduites des personnes amenées de Paris par autobus.

Le champ de l'enquête de l'Elan retrouvé est le même que celui de l'Ined : Paris intra-muros pour les points repas et pour les personnes hébergées auquel s ajoutent

2 Plus de quinze jours.
3 En collaboration avec le Plan Urbain.

les personnes conduites à partir de Paris dans les centres périphériques de la capitale. L'échantillon représentatif de l'enquête a été constitué à partir de personnes ayant recours à des services : hébergement et repas. Cette description s'applique aux sans abri utilisateurs de services. Différentes sources semblent confirmer que la plupart des personnes sans-abri utilisent l'une ou l'autre des ressources. Un échantillon dit "rue" a été cependant constitué à titre exploratoire pour vérifier cette hypothèse.

Le plan de sondage

L'enquête de l'Ined se déroule le jour, sur un laps de temps d'un mois, auprès des utilisateurs des services du type hébergement, distribution de nourriture et de repas chauds, y compris itinérants.

Il s'agit d'un sondage à deux degrés sur les bénéficiaires de services dans Paris intra-muros[4]. Nous avons dû nous limiter à trois types de services afin d'éviter des erreurs de couverture. Il s'agit de :

- les centres de distribution de repas gratuits et points-soupes dont la prestation est le repas (4309 repas un jour moyen, selon l'estimation retenue lors du tirage).

- les centres d'hébergement d'urgence (dont la prestation est la nuitée d'un adulte) : 2701 pour un jour moyen.

- les CHRS[5] et centres de longue durée (dont la prestation est la nuitée d'un adulte) : 4931 personnes.

Cependant, des tests préliminaires ont montré l'intérêt d'enquêter dans les vestiaires, les centres d'accueil de jour et dans les centres de soins où certaines personnes se présentent qu'on ne retrouve pas ailleurs (personnes hébergées chez un parent, squatters).

La méthode de sondage employée consiste à tirer au hasard un échantillon de prestations individuelles parmi les centres de la zone pendant une période donnée, et à enquêter leur bénéficiaire. Il convient donc de dénombrer ces prestations, de les sonder et d'en déduire les probabilités de tirage induites pour les personnes tirées, en tenant compte de la multiplicité de la base de sondage.

Pour accorder une représentation suffisante des personnes les plus marginalisées, l'Ined a surreprésenté l'hébergement d'urgence (taux 1/192) par rapport aux repas

4 A laquelle s'ajoutent les deux centres d'hébergement de banlieue cités pour lesquels les personnes hébergées sont amenées de Paris intra-muros par un service de transport spécialisé.
5 Centres d'hébergement et de réadaptation sociale.

(1/345ème) et à l'hébergement de longue durée (1/455ème). **L'équipe de l'Elan retrouvé** a opté pour le plan plus simple du sondage uniforme parmi ces prestations.

Les unités primaires du sondage sont les "centres-jours", c'est-à-dire l'ensemble des prestations distribuées par un centre donné tel jour d'enquête[6]. On échantillonne donc les lieux et les jours. Le tirage des centres s'est fait proportionnellement à leur capacité d'accueil hebdomadaire (pour tenir compte des jours de fermeture). Pour chacune des quatre semaines d'enquête, quatre des cinq journées ouvrées ont été tirées au sort et affectées par choix raisonné aux centres tirés. Six centres étaient désignés pour chaque jour d'enquête, soit quatre-vingt-seize unités primaires (centres-jours) correspondant à cinquante-six centres différents.

Pour réduire la variance imputable à l'hétérogénéité des clientèles des centres, on a procédé à un tirage sans remise après une "stratification implicite", les centres d'hébergement étant, pour chacune des deux bases, triés par catégorie de population accueillie[7], puis par taille décroissante. Les centres de repas étaient directement classés par taille.

Dans chacune des deux catégories, les centres ont été sélectionnés proportionnellement à leur capacité théorique. Les lieux de distribution des repas ont ensuite été sélectionnés avec une probabilité proportionnelle au nombre moyen de repas servis par semaine.

Le tirage des unités secondaires (les prestations) s'est fait aléatoirement à raison de six prestations par unité primaire (centre-jour) enquêtée. Ainsi dans chaque base, le sondage des prestations est auto-pondéré. Le passage de la prestation à son bénéficiaire se fait par le calcul de la pondération.

L'Ined a conduit, de façon anonyme, **un total de 591 questionnaires** (sur les 606 souhaités à l'origine:

- 219 questionnaires en hébergement d'urgence ;

- 137 questionnaires dans des CHRS ou autres hébergements de longue durée ;

- 235 questionnaires dans des sites de distribution de repas.

A près concertation avec l'Ined, **l'Elan retrouvé** a constitué un échantillon **de 838 personnes** sur 1416 échantillonées.

---

6 Ainsi aucune unité primaire n'est tirée plusieurs fois alors que certains **centres** le sont à des jours distincts.

7 Centres pour hommes seuls ; pour hommes et femmes ; pour hommes, femmes et couples avec enfants ; pour femmes avec enfants ; pour femmes seules.

Les centres d'hébergement ont été stratifiés par type, mais la frontière entre urgence et hébergement longue durée a reposé sur une durée maximale (ou moyenne) de séjour autorisée de 6 mois ou plus, et non sur le statut (CHRS par exemple). Cette distinction de l'Ined a été conservée pour mettre à jour les deux bases de sondage de l'hébergement avant le tirage. La notion d'hébergement d'urgence regroupe les centres où les séjours durent une nuit ou plusieurs semaines, sans précision possible. On sait en effet que dans beaucoup de centres, en période hivernale, les lits sont gardés plusieurs jours de suite par la même personne, et le renouvellement des personnes se fait de séquences en séquences (3 jours par 3 jours, ou semaine par semaine, ou du jour au lendemain, etc...).

La troisième base de sondage a été constituée des points-repas où, de façon fixe ou itinérante, le matin, à midi ou le soir, sont effectuées des distributions de collations, paniers-repas ou vrais repas.

L'équipe de l'Elan retrouvé a mis à jour la base de données qui avait servi à l'enquête de l'Ined de 1995, puis l'Ined a effectué le tirage selon un taux de sondage unique dans les trois bases.

Après le tirage les listes d'hébergement provisoire et d'hébergement de longue durée ont été regroupées dans la même liste "Hébergement" (462 unités pour 52 centres) à laquelle s'ajoutent les 438 unités de 35 points repas. Il a fallu cependant écarter de l'échantillon 54 occupants de la maison d'accueil de Nanterre domiciliés à Nanterre (et non à Paris) qui pour la plupart font des séjours très longs et ne rentraient pas dans le champ de notre enquête. De plus deux centres d'hébergement n'existaient plus au moment de l'enquête et ont été remplacés par des centres de la même strate ; deux autres centres ont refusé de participer et ont été remplacés suivant les mêmes modalités ; un des centres repas avait disparu et n'a pas été remplacé.

Ainsi, après quelques mises au point de la liste, la collecte a conduit finalement à :

Pour l'hébergement : 462-54 = 408
Pour les Points-Repas : 438 -12 = 426
Total = 834

Quelques erreurs de programmation ou de réalisation ont abouti à quelques questionnaires réalisés en trop dans deux centres d'hébergement, et dans un cas en moins. Un questionnaire a été réalisé en trop dans un Point-repas. Un questionnaire n'a pu être réalisé dans un centre de très petite taille.

Ont été réalisés finalement :

dans les centres d'hébergement tirés au sort : 411 questionnaires.
dans les Point-repas tirés au sort : 427 questionnaires.
Au Total = 838.

Enfin 43 questionnaires ont été réalisés à titre exploratoire dans la rue.

2 c. La collecte

L'établissement de la liste des services de Paris intra-muros

La première étape de cette enquête a été l'établissement difficile d'un liste EXHAUSTIVE (dans la mesure du possible) des services d'hébergement et de distribution de nourriture dans Paris intra-muros.

En comparant les diverses sources utilisées, on constate des similitudes rassurantes mais aussi d'apparentes divergences dans les effectifs, les adresses, la spécificité du service rendu : les différents guides répondent eux-mêmes à des objectifs spécifiques selon leur mission (insertion, urgence...).

L'établissement de cette liste a nécessité deux personnes à temps plein pendant trois mois et une mobilisation de tout le service des enquêtes pour les dernières vérifications. La base de sondage obtenue comportait 36 centres d'hébergement d'urgence, 46 CHRS et centres d'hébergement de longue durée, 58 sites de distribution de nourriture. L'expérience a montré qu'il restait quelques erreurs : centre d'hébergement pour mineurs, association n'assurant pas elle-même l'hébergement.

Le tirage des services enquêtés

Nous avons renoncé à enquêter au moment des petits déjeuners. Toutefois, il arrive que la frontière entre petit déjeuner et repas de midi soit elle aussi floue. Dans certains lieux, par exemple, on peut manger un sandwich entre 9 h et 11 h du matin.

Au total, nous avons retenu 98 unités primaires, soit environ six unités par jour sur 16 jours d'enquête, correspondant à une soixantaine de sites.

Quatre jours ouvrés par semaine ont été tirés pour chacune des quatre semaines de l'enquête, entre le 13 février et le 10 mars. Chaque centre ou lieu de distribution de nourriture a été affecté à un ou plusieurs jours. Nous avons exclu d'enquêter le week-end en raison de pratiques différentes ces jours-là (séjour dans la famille ou

chez des amis, par exemple), contrairement à l'enquête de l'Elan retrouvé qui s'est également déroulée le samedi.

Une lettre a été envoyée aux lieux sélectionnés. Nous n'avons reçu que très peu de réponses, même pour refuser. A posteriori, cela n'est pas vraiment surprenant étant donné la charge de travail des organismes au coeur de l'hiver. Nous avons alors pris contact par téléphone avec tous les services tirés. Cette procédure a été plus lourde que prévue. Les rares sites qui ont refusé l'enquête ont été remplacés par un site de même type, dans la même strate.

En fin de compte, il n'y a eu que cinq refus définitifs. Il est clair que ces refus et ces réticences s'expliquent très bien par le souci louable de protéger les personnes dont on s'occupe, et aussi par une charge de travail importante à cette période de l'année.

La sélection des personnes à interroger

L'équipe de l'Ined a dû visiter chaque site pour mettre au point un protocole de tirage adapté au contexte local et le moins éloigné possible du tirage théorique (prendre une personne sur trois dans une file et enquêter la première qui accepte, puis recommencer, pour les sites sans liste préalable ; sinon,tirage sur liste établie d'avance). Ainsi, dans les points-soupes (distributions par des camions, à l'extérieur, souvent la nuit) les personnes se présentent plutôt par masses compactes et sont rarement rangées en files. Pour nous, l'essentiel était de préserver le caractère aléatoire de l'échantillon.

Pour les CHRS nous avions cherché à tirer les personnes à l'avance sur une liste, puis nous leur avions adressé une lettre (mise dans leur casier) pour les rencontrer ultérieurement. A quelques exceptions près, cette procédure s'est révélée peu efficace, les personnes ne se présentant pas et n'en avisant pas le responsable du centre. Devant ces difficultés, nous avons demandé aux responsables de contacter elles-mêmes les personnes après tirage aléatoire. Les refus étaient alors beaucoup moins nombreux, mais cette méthode allait à l'encontre de nos décisions de départ (souci de neutralité par rapport aux organismes gestionnaires). Outre cette difficulté, nous nous sommes quelquefois trouvés confrontés à un autre problème : il est arrivé que trop de personnes acceptent l'interview (et il fallut alors trouver une solution pour ne pas désappointer les personnes qui avaient accepté de nous aider en répondant au questionnaire).

L'accueil fait à l'enquête par les responsables des services est quelquefois mitigé. En particulier, nous avons été très gênés par le passage peu de temps auparavant d'une autre enquête réalisée par le CSA à la demande de La Croix/ La Rue/ la FNARS, à la suite de laquelle certains responsables ne voulaient plus entendre parler d'enquêtes dans leur centre. De plus, les centres font très souvent l'objet de sollicitations de la part de journalistes. Certains ont même été victimes de caméras cachées. On

comprend donc leur lassitude et leur souci de protéger les personnes démunies qui font appel à eux.

L'accueil fait par les personnes interrogées est quelquefois mauvais, mais lorsque l'entretien est accepté il se déroule très bien. Les enquêtés sont en général satisfaits de la relation qui s'est établie avec l'enquêteur.

Les conditions matérielles de réalisation de l'enquête sont très diverses. Dans certains centres, il y a des difficultés à trouver un endroit un peu retiré pour passer le questionnaire dans de bonnes conditions de confidentialité. Dans d'autres, chaque enquêteur a un bureau à sa disposition. Dans la rue (soupes), il faut souvent aller dans un café proche en raison du froid et de la pluie.

Par ailleurs, il a été jugé très important d'enregistrer les refus et les personnes que l'on n'a pas pu interroger (en raison d'un problème de langue par exemple) sur la feuille de contact prévue à cet effet. Chaque enquêteur avait pour consigne, après sélection aléatoire de la personne à contacter, de noter avec le plus de précision possible l'issue de ce contact. L'enquêteur notait aussi le sexe et l'âge approximatif de la personne refusante.

Signalons encore deux points relatifs à la représentativité de l'échantillon :

- des raisons de coût et de complexité de montage nous ont conduits à ne réaliser les entretiens qu'en français. Il est difficile d'avoir sur chaque site des interprètes pour les langues diverses qu'on peut rencontrer à Paris (langues des pays de l'Est, en particulier) et contrairement aux enquêtes que l'on mène dans un logement, on n'est pas assuré de retrouver les personnes de langue étrangère pour réaliser plus tard l'entretien avec un interprète. Dans ces conditions, l'échantillon ne peut être complètement représentatif.

- dans certains sites où les personnes se déplacent rapidement dans tous les sens et où l'atmosphère est quelquefois tendue, les enquêteurs ont eu des difficultés à appliquer le principe de tirage aléatoire mis au point et à remplir complètement la feuille de contact. Le nombre des contacts et des refus s'en trouve donc sous-estimé, en particulier la forme de refus qui consiste à faire semblant de ne pas entendre l'enquêteur et à passer très vite.

Contacts et refus

Au total 997 contacts ont été répertoriés par les enquêteurs. Le taux de réussite s'établit pour l'Ined à 59% sur l'ensemble des sites, atteint 90% dans les CHRS et les centres d'hébergement de longue durée. Il est de 54% dans les centres d'urgence ainsi que sur les lieux de distribution de nourriture. Les conditions objectives de prise de contact sont, en effet, très précaires dans le cas des points-soupes.

Parmi les contacts répertoriés, 406 personnes (soit 41%) n'ont pas répondu, soit à la suite d'un refus explicite ou non, soit pour des raisons d'inaptitude (langue, alcool, drogue...). Ce taux est élevé ; rappelons toutefois qu'il s'agit d'une enquête non obligatoire et qui était explicitement annoncée comme telle ; par ailleurs, nous avions décidé de faire prendre les contacts par les enquêteurs eux-mêmes et non par les responsables des sites, à la fois pour des raisons éthiques (afin que la personne interrogée se sente complètement libre de répondre ou non) et pour des raisons scientifiques (que la réponse ne soit pas orientée par des renseignements antérieurs fournis aux responsables).

Raisons personnelles et facteurs collectifs de refus

Sur les 353 personnes ayant refusé l'entretien de l'Ined, 107 n'ont pas motivé leur refus (dans de nombreux cas il s'agit de personnes s'étant détournées de l'enquêteur ou ayant feint de ne pas l'entendre), 58 se sont déclarées pressées, 32 n'étaient pas françaises, 18 étaient fatiguées ou malades, 18 avaient, selon elles, déjà répondu. Ce dernier cas peut s'interpréter de diverses façons : la personne a pu en effet être déjà interrogée ; il peut s'agir d'une forme polie de refus.

Il y a moins de femmes (12%) parmi les personnes contactées que parmi les questionnaires achevés (20%). Plus qu'une différence de comportement due au sexe, il faut sans doute y voir le reflet des meilleures conditions d'enquête dans les centres accueillant des femmes ou des couples, conduisant à une acceptation plus fréquente.

A l'issue du passage du questionnaire filtre, 44 personnes ont rempli un questionnaire "logement" (4% des contacts, 7% des questionnaires) et 547 un questionnaire "sans logement" (55% des contacts, 93% des questionnaires).

Enfin, les taux de recouvrement entre différents services montrent l'intérêt d'établir des pondérations complexes afin d'assurer la bonne représentativité de l'échantillon, et la part des utilisateurs des points-soupes et autres lieux de distribution de nourriture qui n'utilisent pas les centres, l'intérêt d'enquêter aussi dans ce type de lieux, même si c'est parfois techniquement plus difficile.

Les difficultés de la collecte et leurs implications statistiques

Plusieurs difficultés rencontrées par l'Ined au cours de la collecte perturbent le modèle théorique, notamment :

- certains centres tirés se sont révélés après coup ne pas correspondre au champ de l'enquête (par exemple, des foyers de travailleurs) ou étaient fermés à la date de l'enquête ; ils ont été remplacés par le centre suivant dans la liste ayant servi pour le tirage ;

- pour certains centres, la taille (le nombre moyen de prestations servies par jour, calculé à partir des prestations servies une semaine) était inexacte : taille théorique ne correspondant pas à la taille réelle observée sur le terrain, taille donnée en nombre de repas servis dans les distributions itinérantes de soupe, alors qu'une même personne se sert plusieurs fois, taille en nombre de lits incluant les enfants mineurs des ménages.

- pour des raisons d'organisation, il a quelquefois été nécessaire de modifier le nombre d'enquêtes réalisées dans une unité primaire (tout en restant autour de six) ;

- dans la partie rétrospective sur l'utilisation des services, certains centres ont été insuffisamment précisés, certaines journées mal remplies, soit par l'enquêté, soit par l'enquêteur ;

Ces perturbations tendent à biaiser l'échantillon de deux façons principales :

- un effet sélectif sur les enquêtés entraîne un biais non mesurable et non correctible, dont on peut toutefois espérer qu'il soit négligeable par rapport à l'incertitude d'échantillonnage ;

- des variations d'effectifs (nombre de prestations, nombre d'enquêtés) par unité primaire perturbent le caractère auto-pondéré du sondage.

Il est cependant facile d'effectuer une repondération de la prestation tirée, et par le fait même de l'enquêté.

Pour évaluer le taux de refus et apprécier l'éventualité de refus liés à d'éventuels problèmes de santé mentale, **l'équipe de l'Elan retrouvé** a élaboré une fiche particulière à ce sujet, le rapport de non entretien : il s'agit d'une page remplie par l'enquêteur à "chaque entretien n'ayant pas abouti". Les informations qu'il contient sont appréciées par l'enquêteur, sans être vérifiées (âge approximatif, sexe, pays d'origine probable). L'enquêteur doit de plus documenter trois types d'information :

- si l'enquêté montrait un signe visible de maladie physique ou d'incapacité (oui-non) et le décrire,

- si l'enquêté montrait des signes évidents d'intoxication par l'alcool ou la drogue (oui-non) et les décrire,

- si l'enquêté montrait des signes de maladies mentales telles que "désorienté, parlant tout seul, trop agressif..." (oui-non) et les décrire.

Les réponses aux deux dernières questions ont été relues pour chaque fiche en triant séparément :

les personnes hors champ (ne parlant pas français, ou se déclarant non SDF),
les personnes ayant déjà répondu,
les refus de répondre donnant comme motif: "pas envie, trop pressé, contre les enquêtes, rien à dire, pas envie de parler", etc...ainsi que toutes formulations ainsi rapportées par l'enquêteur.

Les "refus dits pathologiques" ont été validés par un psychiatre au vu de la description du comportement rapporté par l'enquêteur, en incluant les problèmes d'alcool. Certes, il n'est pas possible d'attribuer à la pathologie mentale le "refus de répondre". Toutefois, certains cas le laissaient soupçonner.

Taux de réponse de l'enquête de l'Elan retrouvé

L'entretien a été proposé à 1416 personnes. Parmi elles, 111 étaient hors champ, c'est à dire non SDF, ne parlant pas français ou incapables de communiquer pour d'autres raisons.

Au total, 578 rapports de non entretien ont été analysés ; en moyenne, 24% d'entre eux semblent en relation avec une pathologie mentale, telle que codée par le psychiatre à la lecture du rapport de l'enquêteur. Cette proportion est probablement sous-estimée car un nombre non négligeable de refus de répondre étaient probablement en rapport avec une pathologie mentale qui n'a pu être identifiée. Parmi les refus pathologiques, 15% sont liés à une intoxication par l'alcool et près de 10% à une maladie physique ou à un handicap visible.

Le taux global de réponse est un peu plus élevé que pour l'expérience de l'Ined. Il ne se situe cependant qu'à 64,2% en raison d'un taux relativement bas dans les sites repas (58,6%) n'offrant que des conditions d'entretien difficiles. Au contraire, dans les centres d'hébergement, le taux de réponse est très voisin de celui obtenu en population générale (71,2%). A noter qu'il est légèrement plus élevé en hébergement de longue durée (73%) qu'en hébergement provisoire (70,60%)

2 d. estimation et pondération

L'une des difficultés de cette méthode multi-base est le risque de duplication des personnes sans domicile dénombrées car un sans abri peut être compté sur plus d'un site. Dans une approche de type recensement, la couverture doit être exhaustive et les doubles-comptes repérés et éliminés. La base d'une enquête par sondage doit répondre au même souci d'exhaustivité, mais les " inscriptions" multiples dans cette base sont acceptables sous la condition d'une repondération adéquate des unités enquêtées. Dans un plan de tirage sans remise, aucun individu ne doit figurer plusieurs fois dans l'échantillon enquêté (exigence sans problème en l'absence de

rémunération de l'enquêté). Lorsque la période d'enquête est étendue, le risque de duplication augmente.

Nous distinguons principalement deux étapes pour traiter de ce problème. La première concerne l'identification des exceptionnels doubles comptes et leur élimination du fichier, par des opérations de tri informatique ou manuel ; il est aussi possible de les identifier par une question directe (par exemple, "avez-vous déjà été interrogé ?") ou de combiner les deux approches. Nous avons adopté la première solution.

La deuxième étape concerne la prise en compte des probabilités différentielles d'inclusion dans l'échantillon, et nécessite de recueillir des informations sur l'usage que font des services les personnes sans domicile interrogées.

La pondération tient compte de plusieurs éléments :

Pondération des prestations

Pour l'Ined, les prestations étaient tirées à trois taux de sondage distincts selon qu'il s'agissait de repas, d'hébergement d'urgence ou de longue durée. La pondération de sondage est l'inverse de ces taux de sondage.

Un facteur correctif a dû être introduit au niveau de l'unité primaire pour compenser la mauvaise déclaration de la fréquentation du centre (capacité théorique inexacte, remplissage partiel du centre le jour d'enquête, ou lorsque l'effectif enquêté a été différent de l'objectif de 6 pour cette unité primaire). Cette correction concerne donc les probabilités de tirage de la prestation au second degré.

L'expérience antérieure de l'Ined a permis à **l'Elan retrouvé** d'éviter ces écueils : les effectifs théoriques ont pu être correctement anticipés et le plan de sondage plus strictement respecté. La seule correction significative à apporter tenait à la surestimation systématique du nombre de repas fournis aux points soupes, évaluée à 1/3. C'est la seule correction apportée aux pondérations de prestations initialement uniformes.

Pondérations des individus

La pondération des individus devra être définie en fonction de l'horizon temporel choisi : le cadre statistique le plus simple consiste à vouloir définir la clientèle des ces prestations pour une journée moyenne. On ignore alors si la population se renouvelle d'un jour au suivant. L'enquêté a pu être sélectionné pour trois prestations possibles le jour de son interview (deux repas et une nuitée). Sa probabilité d'inclusion est, à un faible facteur correctif près, la somme des probabilités de tirage des prestations auxquelles il a eu recours ce jour là. Les

estimations de l'enquête seront ensuite la simple moyenne des 16 estimations quotidiennes correspondant à la durée de la collecte.

Un cadre statistique plus intéressant et plus complexe consiste à s'intéresser au renouvellement de la population au cours de l'hiver ou de l'année. Contrairement aux estimations antérieures, on veille à éviter que le SDF stable (en hébergement de longue durée) pendant les 16 jours de collecte ne compte pas autant que les 16 SDF qui seront apparu juste une journée. C'est la raison d'être du questionnaire rétrospectif sur les prestations récemment utilisées. En dénombrant ces prestations, on estime la probabilité qu'avait l'individu d'être enquêté. Malheureusement, il est apparu prudent de limiter la rétrospective systématique à une semaine, avec seulement quelques questions de jalon à l'horizon du mois ou de l'hiver. Car les oublis dans la rétrospective deviennent nombreux et la fiabilité des déclarations est moins qu'incertaine. Ainsi peut-on envisager une étude de la population sans domicile présente au cours d'une semaine, mais pas au cours de l'hiver.

Une difficulté est à prendre en compte lors du calcul de ces pondérations : lorsque l'enquêté a été interrogé à midi, nous n'avons pas d'information sur l'endroit où il mangera et dormira le soir (de même s'il est interrogé au repas du soir, nous ne savons pas où il dormira). Nous avons donc dû compléter ces informations par une procédure d'imputation, établie à partir des informations sur la semaine écoulée[8].

Dans l'échantillon de **l'Elan retrouvé** des questions sur l'utilisation des ressources la semaine précédente et le jour même tant au niveau du logement que des divers repas permettaient d'évaluer le nombre de personnes qui avaient utilisé les ressources contenues dans la base le même jour et qui devaient être pondérées. Cette pondération a touché 234 sujets soit 27,90 % de l'échantillon qui ont été à des degrés divers multiutilisateurs. Le cas le plus fréquent étant celui des personnes qui prennent leurs deux repas par jour dans des lieux de la base soit 131 personnes puis celui des hébergés ayant pris un repas le même jour dans un lieu de la base soit 40 sujets.

A noter que tous les centres des multiutilisateurs ont du être revus pour pouvoir déterminer si le repas du soir avait été pris dans le lieu même de l'hébergement ou dans un lieu différent puisque la pondération ne devait porter que sur ces derniers.

8 Il conviendrait dans une enquête ultérieure d'interroger les enquêtés sur leurs intentions. L'équipe de l'Elan retrouvé a apporté cette amélioration importante de la méthode.

3. Les questionnaires, les résultats et le bilan des enquêtes

L'entretien de l'Ined débute par un questionnaire "filtre" (questionnaire n°1) précisant la situation de logement des personnes enquêtées. Ensuite, les personnes qui disposent d'un logement répondent au questionnaire n°2, celles qui n'en ont pas au questionnaire n°3. L'ensemble des deux questionnaires posés prend environ 30 minutes.

Les questionnaires 2 et 3 abordent un ensemble de thèmes assez large. Ils ne diffèrent que par leur partie "logement". Ils débutent par les caractéristiques démographiques de l'enquêté, suivies de quelques questions sur l'utilisation qu'il a fait des services au cours de la semaine passée, cette partie nous permettant d'établir des pondérations afin de ne pas donner aux personnes qui utilisent beaucoup ces services une importance trop grande. Les questions sur le logement retracent l'histoire résidentielle, et décrivent le logement actuel pour les personnes "logées". Est abordée ensuite l'histoire familiale et les liens subsistant avec la famille, puis le travail, les diplômes et la profession. Enfin, quelques questions portent sur l'origine des ressources financières, (mais non leur montant).

Seule la partie rétrospective du questionnaire sur l'utilisation des services est nécessaire pour établir les pondérations. Les thèmes abordés peuvent être modifiés selon les préoccupations des responsables des enquêtes futures.

Les premiers résultats de l'enquête de l'Ined (Marpsat et Firdion, 1996) font apparaître que les processus conduisant à devenir sans domicile peuvent remonter très loin dans la vie des personnes concernées et qu'une politique de prévention doit tenir compte, non seulement des difficultés de maintien dans le logement et d'accès au logement, mais aussi de la lutte contre tous les aspects de la pauvreté.

Environ un homme sans domicile sur dix avait, en effet, perdu son père avant l'âge de seize ans, autant avaient perdu leur mère. Environ un sur quatre ne vivait à seize ans ni avec son père, ni avec sa mère. Cette proportion est beaucoup plus forte pour les plus jeunes, de même que la proportion de ceux qui étaient à 16 ans en structures collectives (foyers...) ou en famille d'accueil. Presque un homme sans domicile sur cinq ne peut préciser le métier de son père, soit qu'il ne l'ait pas connu, soit que les liens avec lui aient été rompus très tôt.

Au moment de l'enquête, environ un homme sans domicile sur quatre déclare travailler. Mais ce travail est souvent précaire : il ne correspond à un contrat à durée indéterminée (CDI) que dans 17 % de ces cas. Un emploi sur quatre est à durée déterminée (CDD) ou un intérim, les autres étant des CES, d'autres formes d'emploi aidés, ou des « petits boulots ». Près de huit hommes sur dix ont travaillé avant cet

emploi actuel. Pour cet emploi précédent, la précarité était moins grande mais déjà élevée : un tiers seulement des emplois étaient en CDI tandis que 37 % des enquêtés se trouvaient en emploi intérimaire ou en CDD (contre 5 % des hommes actifs occupés pour la France entière, en mars 1995).

Bilan de l'Ined

Ce type d'enquête permet de bien connaître la population sans domicile à un moment donné, y compris dans son parcours. En revanche, elle ne permet pas de connaître les entrées dans et les sorties de la situation de sans domicile au cours d'une longue période, par exemple une année. En particulier, une personne qui a retrouvé un logement ne figure pas dans l'échantillon. Une autre limitation de notre enquête est celle due au champ, mais dans l'état actuel de la connaissance, ce type d'enquête est le meilleur outil même si une partie de la population concernée y échappe.

Pour les enquêtes futures, les chercheurs de l'Ined suggèrent les recommandations suivantes :

À reproduire ? Dans quelles conditions ? A quel coût ?

- A reproduire à l'échelle d'une agglomération ou d'un ensemble de villes de plus de n habitants (par exemple plus de 100 000 ou 150 000 h.), y compris comme complément d'une enquête auprès de la population logée (enquête sur l'emploi, sur la santé). Dans ce cas il conviendrait que soit réalisé en amont un inventaire des services existants propre à une utilisation statistique, que chaque gestionnaire d'enquête n'aurait plus qu'à réactualiser.

- En se reposant sur des relais associatifs et institutionnels locaux, le réseau INSEE dans le cas français, et les milieux universitaires de chaque ville.

- En planifiant l'établissement de la liste des services, la confection du questionnaire, les contacts avec les services ; les tests peuvent se faire la première année, durant la constitution de la base de sondage, l'enquête proprement dite la seconde année.

- Reproduire l'enquête à des saisons différentes avec le même type de conception d'échantillon aléatoire.

- Le coût externe a été, pour l'opération parisienne, de l'ordre de 400 000F (sans compter les opérations méthodologiques spécifiques qui ne sont pas à reproduire à grande échelle) pour 600 enquêtes réalisées (667 F par questionnaire).

Avec quels aménagements ?

• nous ne recommandons pas le recensement de rue durant la nuit (coûteux et peu productif) bien que cela permette d'interroger quelques personnes sans domicile ne faisant appel à aucun service.

• en remplacement de ce dispositif de nuit, nous proposons de compléter l'échantillon "services" par un échantillon de lieux d'enquête de jour dans la rue, repérés comme étant fréquentés de jour par les sans-abri. Les études qualitatives en cours actuellement, ainsi que la collaboration des associations, pourraient nous aider à définir ces lieux. D'autre part, il faudra examiner aussi la possibilité d'utiliser les données du Samu social dans les lieux où un tel dispositif existe.

• il conviendrait de mieux prendre en compte les non francophones.

• se pose le problème d'étendre le champ des services au-delà de l'hébergement et de la restauration. Mais il faut prendre garde dans ce cas de ne pas introduire de défaut d'exhaustivité.

• il faudra se poser la question de la spécificité du rural.

Questionnaire, résultats et bilan de l'Elan retrouvé

Le questionnaire de l'Elan retrouvé se divise en trois parties :

une première partie décrit l'histoire de l'itinérance et aborde les questions socio-démographiques : formation, histoire professionnelle (nombre, nature et durée des emplois), statut matrimonial, ressources actuelles, lieu de naissance, situation géographique et habitat avant la période d'itinérance, affiliation à la sécurité sociale ;

la seconde partie concerne les maladies physiques et handicaps, et leur prise en charge, dont l'hospitalisation ;

la troisième partie traite des problèmes de santé mentale que l'on souhaitait explorer dans cette population particulière :

- les délires chroniques et les schizophrénies,

- les troubles de l'humeur : manie et troubles dépressifs majeurs psychotiques ou névrotiques y compris les conduites suicidaires,

- les problèmes d'utilisation des toxiques: alcool, drogues ;

- après chacun de ces diagnostics suivent des questions sur l'utilisation du système de soins (hospitaliers ou ambulatoires) et sur la prise de médicaments psychotropes ;

- les problèmes de personnalité : personnalité paranoïaque, personnalité labile, impulsive et personnalité limite.

Les diagnostics psychiatriques reposent sur la stratégie du Diagnostic Interview Schedule (DIS), validée par Robins et Helzer (1981) et reprise par le CIDI (Composite International Diagnostic Interview ; Robins, Wing 1988) . C'est cette méthode qui a été validée avec de bons résultats lors de l'enquête pilote par une centaine de double interviews psychiatres/ interviewers.

La méthode explore les symptômes des différents diagnostics psychiatriques qui sont ensuite organisés suivant les critères des classifications pour aboutir à des diagnostics. Ces symptomes sont exprimés en langage simple et groupés par appartenance diagnostic . Un exemple que l'interviewer doit noter verbatim est demandé en cas d'idées délirantes ; exemples qui sont revus sytématiquement par un psychiatre expérimenté pour validation.

Dans cette enquête le CIDI a été utilisé sous une forme légèrement simplifiée pour les troubles délirants, la manie et les troubles cognitifs (mini mental state). Pour les autres problèmes: troubles dépressifs, troubles de l'usage des toxiques c'est le CIDIS (Composite International Diagnostic Interview Simplified) qui a été choisi . Le CIDIS est une forme encore plus simplifiée et plus rapide que le CIDI dont il utilise néanmoins la plupart des questions ; il avait été utilisé dans l'enquête Santé des Franciliens qui comportait un échantillon de Rmistes domiciliés.

Pour évaluer l'utilisation des services des questions sont posées après chaque diagnostic psychiatrique, qui précisent le type de soignant rencontré (généraliste, psychiatre, psychologue ou autre) et, le cas échéant, la classe du médicament prescrit (anxiolytique, antidépresseur, somnifère), ainsi que son retentissement sur la personne. Le questionnaire permet de connaître les dates du premier et du dernier épisode (Kovess, Chanoit 1992).

Les troubles de la personnalité sont mesurés par un questionnaire présenté sous forme de propositions concernant les attitudes généralement utilisées par la personne. Celle-ci doit répondre par vrai ou faux ; l'instrument s'appelle le PDQ (personality diagnostic questionnaire de Dowson). Traduit et expérimenté en France par Nollet, il permet de sélectionner de 6 à 8 questions par type de personnalité.

Caractéristiques socio-démographiques et biographies

Les caractéristiques socio-démographiques observées concordent pleinement avec celles publiées par l'Ined : ainsi, 60% des sans abri sont nés en France. A contrario, 40% sont nés à l'étranger dont 16% dans le Maghreb et près de 10% en Afrique noire. La très grande majorité de ces personnes a déjà travaillé et souvent pendant plusieurs années. 20% disent avoir eu des postes de cadre et posséder au moins le niveau BAC. 70 % habitaient un appartement, 12 % étaient à l'hôtel. 73 % sont affiliées à la sécurité sociale, avec ou sans la carte Paris Santé. Sur un an, 30 % ont passé une nuit à l'hôpital. 17 % d'entre elles ont été placées dans leur enfance. 22 % ont fait un stage d'insertion.

Cette population reste, malgré les aides proposées, très démunie et un tiers d'entre elle déclare n'avoir aucun soutien ; a contrario 31,7% touche le RMI, 11,2% le chomage et 15,6% un salaire.

Les femmes ne représentent que 14 % de l'échantillon et sont plus nombreuses en hébergement long, et plus rares dans la rue. Elles sont itinérantes depuis plus longtemps, ont plus souvent fait un stage d'insertion ; plus souvent affiliés à la sécurité sociale, elles ont travaillé moins longtemps et davantage comme employées que comme ouvrières. Elles invoquent pour cause de leur situation une rupture familiale plus fréquemment que les hommes qui citent surtout des causes matérielles (perte d'emploi, manque d'argent). Les causes de type psychologique sont aussi très fréquentes.

Premières conclusions de l'enquête de l'Elan retrouvé

La présentation des résultats cliniques nécessiterait un long développement. Ainsi, la régression logistique montre que seul le placement dans l'enfance multiplie le risque de troubles psychotiques (d'un facteur 1,68) .

Les premières conclusions qui se dégagent de ce travail montrent à la fois l'homogénéité frappante de la distribution des problèmes de santé mentale dans une population très hétérogène et en même temps l'importance de sous-groupes spécifiques qu'il convient de traiter comme tels.

Ainsi, globalement à un moment donné (les six derniers mois), 6 % environ de cette population présente un trouble psychotique actif (auxquels s'ajoute une partie des 5 % de troubles psychotiques possibles), on doit ajouter 1,70 % de troubles affectifs bipolaires et 20 % de troubles dépressifs sévères. Les troubles dus à l'alcool concernent 15 % de cette population et 30 % de la population des moins de 30 ans sont concernés par les troubles de l'usage des drogues qui, en moyenne, atteignent 10 % des personnes. Etant donnée la présence simultanée de plusieurs problèmes,

les problèmes que nous avons évalués concernent **29 %** de la population, si l'on ne tient pas compte des troubles de la personnalité.

Qui plus est un certain nombre de problèmes n'ont arbitrairement pas été pris en compte, en particulier les troubles de l'anxiété et la névrose post-traumatique qui, dans certaines études, semble fréquente dans cette population. Les chiffres avancés ici sous-estiment donc plus qu'ils ne surestiment les problèmes.

Ces prévalences, à un moment donné, sont bien entendu inférieures aux prévalences sur la vie : 16 % de troubles psychotiques, 41 % de troubles de l'humeur et 34 % de troubles dus à l'usage de substances toxiques. Au total et en tenant compte de la comorbidité **58 %** des personnes sans-abri ont présenté à un moment ou à un autre un problème psychiatrique ; ce chiffre devenant 68 % si on inclut les troubles de personnalité.

De plus, ces prévalences de troubles psychotiques sur la vie ne tiennent compte que des troubles qui ont été confirmés et n'intègrent pas les 12,20 % de troubles possibles, dont certains seraient très probablement confirmés par un entretien clinique, et pour lesquels les traits de personnalité paranoïaque sont présents dans près de 70 % des cas. Sans que l'on puisse dire qu'il s'agit de personnes présentant des troubles de type schizophréniques ou délirants, on peut faire l'hypothèse qu'il s'agit de personnes présentant de problèmes de santé mentale relativement sévères qui les handicapent dans leur fonctionnement social et peuvent rendre leur réinsertion difficile.

Ces différents problèmes sont présents avec la même fréquence quels que soient le sexe, l'âge, le lieu de naissance et ce que nous avons appelé le lieu de vie (hébergement long terme, provisoire ou rue), sauf en ce qui concerne les troubles de l'usage des substances (alcool et drogues), qui au contraire sont modulés par les facteurs : sexe et lieu de naissance pour l'alcool, et âge pour la drogue. Mais surtout ils sont plus fréquents dans les sous-populations particulièrement marginalisées où la fréquence des troubles de l'usage de l'alcool est multipliée par deux.

On constate le rôle joué par les troubles d'abus de substances, drogues pour les jeunes et alcool pour tous les âges, dans le maintien et l'aggravation de la marginalisation.

Simultanément différents sous groupes méritent d'être individualisés :

- les sans-domicile de moins de 25 ans sont une population très touchée. On peut tenter d'en tracer un tableau de la manière suivante : le tiers d'entre eux ont été placés dans leur enfance et n'ont par conséquent aucune ressource familiale. Ils présentent une intoxication aux drogues sévères, des troubles de la personnalité de type labile (impulsif ou limite) et n'ont pas de ressources.

- les femmes, beaucoup moins nombreuses que les hommes, sont plus facilement "institutionnalisées" soit en centre longue durée pour une réinsertion soit en hôpital psychiatrique en cas de problèmes de ce type ; elles souffrent beaucoup moins de problèmes d'alcool et de troubles de personnalité que les hommes.

- les plus de 55 ans ne sont pratiquement jamais dans des centres de longue durée et ont peu de problèmes d'abus de drogues ; ils présentent des prévalences moins élevées de problèmes de santé mentale. Cependant les différences ne sont pas significatives, sauf pour les troubles bipolaires ,absents dans cette catégorie d'âge.

Bien entendu, ceci ne doit pas faire oublier que le groupe numériquement le plus important est celui des hommes d'âge moyen (25/55 ans) qui utilisent les ressources d'hébergement provisoires et constituent la majorité des personnes les plus marginalisées ayant des problèmes avec l'alcool.

Comme l'ont montré un certain nombre d'études, la population sans-abri est en contact avec le système de soins : 54,50 % des personnes qui présentent un problème de santé mentale dans les six derniers mois a consulté un médecin, 29,6 % un psychiatre ; 26,6 % a été hospitalisé en hôpital général dont 9,4 %, en hôpital psychiatrique. Enfin, 45,5 % a pris un médicament psychotrope. Bien entendu, ces données, issues d'une enquête transversale, ne renseignent pas sur la qualité du suivi qui semble très problématique et pourtant indispensable pour des problèmes de ce type.

Il faut souligner enfin la prudence devant entourer les conclusions des études épidémiologiques de ce type ; malgré les précautions méthodologiques qui sont exposées, un certain nombre de personnes n'ont pu être abordées du fait des refus ou du non recours aux services retenus pour le sondage. Cependant les rapports de "non-entretien" nous renseignent relativement sur l'état des non répondants qui ne semble pas très différent en terme de pathologie.

De plus l'analyse des resultats de l'échantillon de "Rue" montre qu'il se rapproche d'une partie de l'échantillon enquêté dans les points repas et que nous avions qualifié de « marginaux » à partir d'une déclaration de plus de 5 nuits passés dans des lieux de type métro, cartons , parkings soit 13,6% de l 'échantillon. En fait l'échantillon « rue » est moins marginalisé que ce sous groupe en ce sens que la moitié seulement des personnes ainsi interrogées ont passés plus de 5 nuits dans ce type de lieu l'autre moitié ayant utilisé les hébergements ou dormi chez des amis ou à l'hotel. Les deux groupes présentent les mêmes types de pathologie.

Bibliographie

Une bibliographie détaillée est disponible auprès des auteurs.

BURT Martha R., 1992, *Practical Methods for Counting Homeless People*, Washington : Interagency Council for the Homeless and Department of Housing and Urban Development.

CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE, 1996, *Pour une meilleure connaissance statistique des sans-abri et de l'exclusion du logement*, rapport final, Paris, n°29, mars 1996 et Actualités du CNIS n° 17, mai 1996. (Version anglaise : *Towards a better understanding of the homeless and exclusion from housing*).

DENNIS Michael L. et IACHAN Ronaldo, 1993, « A Multiple Frame Approach to Sampling the Homeless and Transient Population », *Journal of Official Statistics*, 9(4).

FIRDION J.-M. et MARPSAT, M., 1994, « La statistique des sans domicile aux Etats-Unis », *Courrier des Statistiques*, n°71-72, décembre 1994.

KOVESS V., CHANOIT P.F., DE VIGAN C., « Le CIDISA, une méthode rapide de détection des diagnostics psychiatriques : résultats d' une enquête préliminaire dans les Yvelines », *L'Evolution psychiatrique*, 1992, n°57, 2, pp 225 à 236.2.

KOVESS V., GYSENS S., POINSARD R., CHANOIT P.F., « La psychiatrie face aux problèmes sociaux : la prise en charge des RMIstes à Paris », *L'Information psychiatrique*, 1995, 3, pp 273-285.

MARPSAT M. et FIRDION J. M., « Devenir sans-domicile : ni fatalité, ni hasard », Ined, *Population & Sociétés*, n°313, mai 1996. (Version anglaise disponible : « Becoming homeless : who is at risk ? »).

MARPSAT M. et FIRDION J. M., « Les sans-domicile à Paris : une enquête sur échantillon représentatif de la clientèle des services aux sans domicile » Communication au séminaire de la Feantsa à Vienne, juillet 1996, Ined (version anglaise disponible).

Le rapport de l'enquête de l'Elan retrouvé est disponible sur demande 23 rue de Larochefoucauld 75009.

INSEE

L'INFORMATION SUR L'INFORMATION

INSEE ACTUALITES

"INSEE ACTUALITÉS magazine" est un catalogue trimestriel des nouveautes de l'INSEE : publications, banques de données... ; il est adressé à toute personne ou organisme désireux de suivre l'actualité de l'INSEE.

Abonnement gratuit sur simple demande à :
Insee - Direction générale
Abonnement à Insee Actualités - Timbre H533
18 bd A. Pinard - 75675 Paris cedex 14

COURRIER DES STATISTIQUES

Quatre fois par an cette revue interministérielle vous informe sur l'ensemble des activités du système statistique public et sur l'évolution des outils et des méthodes.

Abonnement 1 an (4 numéros)
France : 135 FF - Europe : 169 FF - Reste du monde : 234 FF

BLOC-NOTES DE INSEE INFO SERVICE

A la fois un répertoire et un guide de l'information économique. Le "thème du mois" fournit des repères sur un sujet d'actualité.

Abonnement 1 an (11 numéros)
France : 185 FF - Europe : 231 FF - Reste du monde : 351 FF

SCRIBECO

Une revue bibliographique reflet du fonds documentaire de l'INSEE.

Abonnement 1 an (6 numéros)
France : 657 FF - Europe : 821 FF - Reste du monde : 892 FF

LES PÉRIODIQUES

LE BULLETIN MENSUEL DE STATISTIQUE

10 000 séries mensuelles, trimestrielles et annuelles concernant l'ensemble de la vie économique, complétées par les séries rétrospectives des principaux indices et par le bilan démographique.

Abonnement 1 an (12 numéros)
France : 364 FF - Europe : 455 FF - Reste du monde : 584 FF

INSEE PREMIÈRE

Le "4 pages" qui, chaque semaine, présente les analyses et les commentaires des experts de l'INSEE sur un thème de l'actualité économique et sociale.

Abonnement (60 numéros)
France : 530 FF - Europe : 663 FF - Reste du monde : 827 FF

ÉCONOMIE ET STATISTIQUE

Chaque numéro est un recueil d'articles sur un grand thème du débat social proposant des commentaires, des tableaux et des graphiques ainsi qu'une bibliographie.

Abonnement 1 an (10 numéros)
France : 414 FF - Europe : 518 FF - Reste du monde : 633 FF

ANNALES D'ÉCONOMIE ET DE STATISTIQUE

Ce trimestriel publie des travaux originaux de recherche théorique ou appliquée dans les domaines de l'économie, de l'économétrie et de la statistique.

Abonnement 1 an (4 numéros)
France : 470 FF - Europe : 588 FF - Reste du monde : 629 FF
Pour les particuliers :
France : 170 FF - Europe : 212 FF - Reste du monde : 253 FF

INSEE MÉTHODES

La méthodologie des travaux de l'INSEE et les modèles.

Abonnement (10 numéros)
France : 728 FF - Europe : 910 FF - Reste du monde : 1 103 FF

INSEE RÉSULTATS

Cette série présente les résultats détaillés des enquêtes et opérations statistiques menées par l'INSEE.
Elle s'articule en 5 thèmes :

Économie générale (20 numéros)
France : 1 454 FF - Europe : 1 818 FF - Reste du monde : 2 075 FF

Démographie - Société (7 numéros)
France : 509 FF - Europe : 636 FF - Reste du monde : 726 FF

Consommation - Modes de vie (10 numéros)
France : 728 FF - Europe : 910 FF - Reste du monde : 1 050 FF

Système productif (15 numéros)
France : 1 091 FF - Europe : 1 364 FF - Reste du monde : 1 557 FF

Emploi - Revenus (18 numéros)
France : 1308 FF - Europe : 1 635 FF - Reste du monde : 1 860 FF

Ensemble des 5 thèmes (70 numéros)
France : 5 090 FF - Europe : 6 363 FF - Reste du monde : 7 259 FF

INSEE

COLLECTION SYSTÈME STATISTIQUE PUBLIC

RECUEIL D'ÉTUDES SOCIALES

Une sélection d'études sur l'actualité sociale les plus récemment publiées par différents organismes publics français d'études et de statistiques.

Abonnement 1 an (3 numéros)
France : 315 FF - Europe : 394 FF - Reste du monde : 421 FF

SYNTHÈSES

Cette nouvelle collection présente des études et des enquêtes faites par les organismes du système statistique public.

Abonnement 1 an (6 numéros)
France : 436 FF - Europe : 545 FF - Reste du monde : 623 FF

LA CONJONCTURE
COLLECTION "INSEE CONJONCTURE"

INFORMATIONS RAPIDES

Série de 350 numéros par an, présentant dès leur disponibilité les derniers indices et les résultats les plus récents des enquêtes de conjoncture de l'INSEE.

Elle inclut les 105 numéros des "Principaux indicateurs" (chiffres essentiels de l'économie) qui peuvent faire l'objet d'un abonnement à part par courrier ou par télécopie.

Abonnement
Principaux Indicateurs (105 numéros par an):
. par télécopie :
France : 2 000 FF - Europe : 2 500 FF - Reste du monde : 3 000 FF
. par courrier :
France : 830 FF - Europe : 1 038 FF - Reste du monde : 1 288 FF

Abonnement Informations Rapides (245 numéros par courrier) + les principaux Indicateurs par télécopie :
France : 2 950 FF - Europe : 3 688 FF - Reste du monde : 4 563 FF

Abonnement à l'ensemble de la série par courrier :
France : 1 750 FF - Europe : 2 188 FF - Reste du monde : 2 691 FF

TABLEAU DE BORD HEBDOMADAIRE

Un panorama complet et actualisé de la conjoncture économique française et internationale. Le supplément "Série longues " donne des tableaux et des graphiques sur 25 ans.

Abonnement 1 an (50 numéros + 1 supplément Séries longues)
France : 1 500 FF - Europe : 1 875 FF - Reste du monde : 2375 FF

NOTE DE CONJONCTURE

Trois notes de synthèse et un point de conjoncture pour suivre la situation et les perspectives à moyen terme de l'économie française. Le supplément "Séries longues" donne des tableaux et des graphiques sur 25 ans.

Abonnement 1 an (3 notes + 1 point + 1 supplément Séries longues)
France : 200 FF - Europe : 250 FF - Reste du monde : 296 FF

CONJONCTURE IN FRANCE

Deux fois par an une synthèse de la conjoncture économique de la France rédigée en anglais.

Abonnement 1 an (2 numéros)
France : 48 FF - Europe : 60 FF - Reste du monde : 72 FF

NOTE DE CONJONCTURE INTERNATIONALE DIRECTION DE LA PRÉVISION

Deux fois par an, un panorama de la conjoncture mondiale dressé par la Direction de la Prévision. En supplément, deux points de conjoncture internationale.

Abonnement 1 an (2 notes + 2 points)
France : 150 FF - Europe : 188 FF - Reste du monde : 239 FF

BULLETIN D'ABONNEMENT

À RETOURNER À : INSEE - CNGP BP 2718 - 80027 AMIENS Cedex 01
Tél. : 03 22 92 73 22 - Fax : 03 22 97 92 95

Veuillez noter mon abonnement aux publications suivantes : .....
.....
.....

Nom ou raison sociale : .....

Activité : ..... Tél : ..... Fax : .....

Adresse : .....
.....

Je règle un montant de ..... FF* (total des abonnements) par : ☐ chèque (à l'ordre de l'Insee).

☐ Carte bancaire. ☐ Visa ☐ Mastercard ☐ Eurocard *(seules cartes acceptées)*

Carte N° |_|_|_|_| |_|_|_|_| |_|_|_|_| |_|_|_|_| Expire au : |_|_| |_|_|

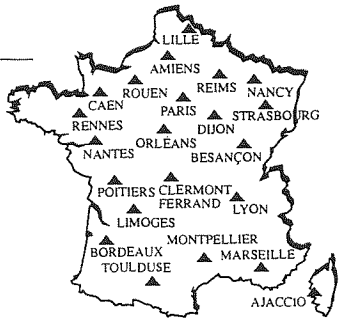
Date : Signature obligatoire :

**pour l'Europe libellé en FF.*
N.B. : Toute commande par fax devra être obligatoirement confirmée par courrier.

L'INSEE DANS VOTRE RÉGION

VOUS Y TROUVEREZ :

- Salle de documentation en libre consultation
- Bureau de vente des publications de l'INSEE
- Adresses des entreprises et établissements (SIRENE).
- Accès au fonds documentaire et aux banques de données de l'INSEE.
- Travaux à la demande...



LE SERVICE INSEE 24H/24

08 36 68 07 60 (2.23 F/mn)

- **indices**
- **informations**
- **adresses**

et sur Minitel

36.15 INSEE (1,01F/mn)
36.17 INSEE *les informations directement chez vous par télécopie* (5,57F/mn)

ALSACE
Cité administrative,
2, rue de l'Hôpital Militaire,
67084 STRASBOURG CEDEX
Tél. : 03 88 52 40 43

AQUITAINE
33, rue de Saget,
33076 BORDEAUX CEDEX
Tél. : 05 57 95 04 00

AUVERGNE
3, place Charles de Gaulle, BP 120,
63403 CHAMALIERES CEDEX
Tél. : 04 73 31 82 00

BOURGOGNE
2, rue Hoche, BP 1509,
21035 DIJON CEDEX
Tél. : 03 80 40 67 48

BRETAGNE
"Le Colbert",
36 place du Colombier,
35082 RENNES CEDEX
Tél. : 02 99 29 33 33

CENTRE
43, avenue de Paris, BP 6719,
45067 ORLÉANS CEDEX 2
Tél. : 02 38 69 53 35

CHAMPAGNE-ARDENNE
1, rue de l'Arbalète,
51079 REIMS CEDEX
Tél. : 03 26 48 61 00

CORSE
Résidence Cardo,
rue des Magnolias,
BP 907,
20700 AJACCIO CEDEX 9
Tél. : 04 95 23 54 50

FRANCHE-COMTÉ
Immeuble "Le Major",
83, rue de Dôle,
BP 1997,
25020 BESANCON CEDEX
Tél. : 03 81 41 61 66

ILE-DE-FRANCE
INSEE Info Service,
accueil, librairie, consultation,
travaux sur mesure et sur rendez-vous
Tour "Gamma A",
195, rue de Bercy,
75582 PARIS CEDEX 12
Tél. : 01 41 17 66 11

Direction Régionale
7, rue Stephenson,
Montigny-le-Bretonneux
78188 ST-QUENTIN-EN-YVELINES CEDEX
Tél. : 01 30 96 90 99

LANGUEDOC-ROUSSILLON
274, allée Henri II de Montmorency,
"Le Polygone",
34064 MONTPELLIER CEDEX 2
Tél. : 04 67 15 71 11

LIMOUSIN
50, avenue Garibaldi,
87031 LIMOGES CEDEX
Tél. : 05 55 45 20 07

LORRAINE
15, rue du Général Hulot, BP 3846,
54029 NANCY CEDEX
Tél. : 03 83 91 85 85

MIDI-PYRÉNÉES
36, rue des 36 ponts,
31054 TOULOUSE CEDEX
Tél. : 05 61 36 61 13

NORD - PAS-DE-CALAIS
130, avenue du Président J.-F. Kennedy,
BP 769, 59034 LILLE CEDEX
Tél. : 03 20 62 86 33

BASSE-NORMANDIE
93-95, rue de Géôle,
14052 CAEN CEDEX
Tél. : 02 31 15 11 11

HAUTE-NORMANDIE
8, quai de la Bourse,
76037 ROUEN CEDEX
Tél. : 02 35 52 49 94

PAYS DE LA LOIRE
105, rue des Français Libres, BP 67401,
44274 NANTES CEDEX 02
Tél. : 02 40 41 79 80

PICARDIE
1, rue Vincent Auriol,
80040 AMIENS CEDEX 1
Tél. : 03 22 91 39 39

POITOU-CHARENTES
5, rue Sainte Catherine, BP 557
86020 POITIERS CEDEX
Tél. : 05 49 30 01 01

PROVENCE - ALPES - CÔTE D'AZUR
17, rue Menpenti,
13387 MARSEILLE CEDEX 10
Tél. : 04 91 17 59 60

RHÔNE-ALPES
165, rue Garibaldi, BP 3196,
69401 LYON CEDEX 03,
(Cité administrative de la Part-Dieu)
Tél. : 04 78 63 22 02

EN OUTRE - MER :

ANTILLES-GUYANE
Direction Inter-Régionale
Tour Secid, 7ème étage,
Place de la rénovation, BP 300
97158 POINTE-A-PITRE CEDEX
Tél. : 05 90 91 59 80

GUADELOUPE
Service Régional
Rue Paul Lacavé, BP 96,
97102 BASSE-TERRE
Tél. : 05 90 99 36 36

GUYANE
Service Régional
1, rue Maillard Dumesle, BP 6017,
97306 CAYENNE CEDEX
Tél. : 05 94 31 61 00

MARTINIQUE
Service Régional, Centre Delgrès
Boulevard de la Pointe des Sables
Les Hauts de Dillon, BP 641
97262 FORT DE FRANCE CEDEX
Tél. : 05 96 60 73 60

RÉUNION
Direction Régionale,
15, rue de l'Ecole, BP 13,
97408 ST DENIS MESSAG CEDEX 9
Tél. : 02 62 48 89 21

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

INSEE - DIRECTION GÉNÉRALE
Unité Communication Externe
Timbre H501 - 18, bd Adolphe-Pinard
75675 Paris Cedex 14 - FRANCE

Tél. renseignements : 01 41 17 66 11
Tél. administration : 01 41 17 50 50

INSEE MÉTHODES
N° 69 - 70 - 71

INSEE MÉTHODES

ACTES DES JOURNÉES DE MÉTHODOLOGIE STATISTIQUE

11 et 12 décembre 1996

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Ce volume rassemble les communications des V-èmes " Journées de méthodologie statistique " qui se sont tenues à Paris les 11 et 12 décembre 1996.

Les thèmes abordés sont les questionnaires et réponses aux enquêtes, les séries temporelles, les mesures d'inégalité, les statistiques locales et les enquêtes sur des sujets sensibles.

Ces journées poursuivent un double but :

- présenter des travaux actuels réalisés à l'Insee à un large public ;

- bénéficier du regard critique d'experts venus de l'étranger qui, en retour, présentent leurs travaux.

On appréciera donc spécialement les interventions des Canadiens Jean-René Boudreau sur le problème de la confidentialité des données et Jean-François Gosselin sur la pratique des enquêtes par téléphone à Statistique Canada.

ISSN 1142 - 3080
ISBN 2-11-066600-5
IMETO69
Août 1997 - Prix : 228 F