Colloque francophone sur les sondages 2005

PREMIERS SONDAGES FRANÇAIS DANS LES DOSSIERS DE SÉCURITÉ SOCIALE ET APPARIEMENT AVEC LES ENQUÊTES AUPRÈS DES MÉNAGES

Andrée MIZRAHI[1] et Arié MIZRAHI[1]

Résumé : C'est l'histoire qui a conduit en 40 ans d'une réflexion sur les méthodes d'enquête et d'une utilisation novatrice des informations figurant dans les dossiers de Sécurité sociale au panel national sur les soins et la santé. Cette méthodologie a été reprise dans d'autres domaines.

Dans les années 60, les statistiques en micro-économie de la santé étaient issues d'enquêtes auprès des ménages ou des producteurs de soins. Entre 1965 et 1975, l'utilisation des dossiers de Sécurité sociale est envisagée pour estimer la consommation et les remboursements, étudier les relations avec la morbidité et les variables socio-démographiques, tester l'effet de décisions à prendre et analyser les données longitudinales.

Le projet de sondage aléatoire dans les dossiers de Sécurité sociale est explicité en 1975, la méthode élaborée en 1977 et validée en 1978. Le sondage, de 4 départements en 1978, est devenu national en 1989 : 11 ans de montée en charge pour le Régime général. Le régime des indépendant a participé en 1994 et le régime agricole en 1996. Le taux de sondage a été doublé (1/600) en 2000.

Après une enquête pilote en 1986-87, l'enquête complémentaire auprès des ménages (prévue dès l'origine), a démarré en 1988. L'enquête, annuelle au départ, est passée à un rythme bisannuel. L'appariement est fait en routine depuis 2000.

Les fichiers de Sécurité sociale sont adaptés à la gestion des remboursements, les transformer en fichiers utilisables pour la recherche et l'appariement avec une enquête ménage demandent un investissement important.

* ** *** ** *

> On peut faire des sondages dans des masses de documents existants, par exemple : les dossiers de Sécurité sociale ou les dossiers de malades hospitalisés. L'avantage est le moindre coût : les « questionnaires » existant déjà. L'inconvénient est que ces « questionnaires » ne sont pas conçus à priori pour répondre rigoureusement au problème posé et ne sont pas toujours « renseignés » sur la totalité des questions posées, d'où, souvent, un gros déchet de dossiers inexploitables.
>
> G. Rösch et la Division d'économie médicale du CREDOC, (1973), *Economique médicale*, Paris : Flammarion médecine – sciences.

1. INTRODUCTION

Dans les années 60, les statistiques en micro-économie de la santé étaient issues d'enquêtes auprès des ménages ou auprès des producteurs de soins. Les premiers à utiliser l'information produite par l'activité administrative de la Sécurité sociale ont été les médecins, habitués à travailler sur les dossiers des malades : il s'agissait de la confrontation des données du service médical de la Sécurité sociale et des données de remboursement. Dans un deuxième temps, les données de Sécurité sociale ont permis de conforter les données d'enquête auprès des ménages ; enfin, l'analyse a porté en premier sur les données de issues du versement des prestations.

Avant de présenter les sondages successifs dans les dossiers de Sécurité sociale, il est nécessaire d'en connaître le fonctionnement en relation avec la législation et la réglementation.

1. LA PROTECTION MALADIE EN FRANCE.

La protection maladie est très morcelée en France, horizontalement et verticalement.

---

1 *ARgSES, Arguments socio-économiques pour la santé*

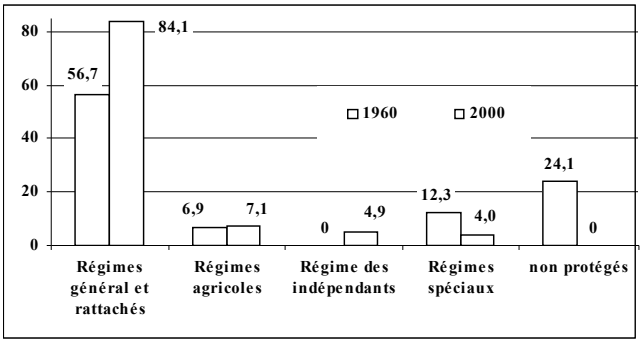
Créée en 1945, héritière des mutuelles, pour la plupart professionnelles, la protection de base était et est encore ventilée en plusieurs régimes selon le statut professionnel : régimes général (salariés du commerce et de l'industrie) et rattachés (fonctionnaires, étudiants, artistes et auteurs, médecins et auxiliaires conventionnés, CMU,…), régimes des indépendants (artisans, commerçants, libéraux), des agricoles (exploitants et salariés), des régimes dits spéciaux (SNCF, EDF-GDF, mineurs, marins pêcheurs, militaires,…) ; chacun de ces régimes avait, et a encore ses règles et ses processus de remboursement. Avec cette logique, seuls les actifs sont assurés (ils paient des cotisations), les personnes à leur charge (conjoints inactifs et enfants) sont leurs ayants droit. De même, toutes les personnes résidant en France n'étaient pas protégées en 1945. Les extensions successives ont conduit à une couverture générale, depuis la loi sur la couverture maladie universelle (CMU, loi du 27 juillet 1999).

Le principal régime (régime général) est géré par 122 caisses primaires d'assurance maladie (CPAM) autonomes, avec conseil d'administration paritaire élu ; au départ, les processus de remboursement n'étaient unifiées (et donc l'organisation de ses fichiers), non plus que les nomenclatures.

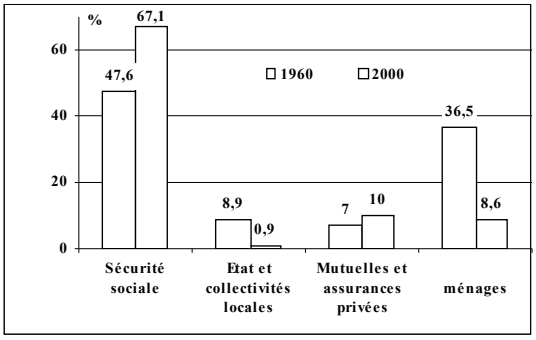
Ces CPAM étaient fédérées dans un organisme national la Fédération nationale des organismes de Sécurité sociale (FNOSS). La FNOSS a commencé un travail d'unification jusqu'à sa dissolution, en 1968[2], remplacée par la caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS), moins indépendante et dont il nomme le directeur. La CNAMTS a poursuivi l'unification des règles de remboursement, des procédures et des nomenclatures, en particulier avec les deux autres grands régimes de Sécurité sociale (indépendants et agricoles).

La protection de base (par les régimes obligatoires) assure un remboursement des dépenses de soins médicaux variable selon différents critères liés à la personne, à la maladie ou à la nature des soins ; globalement, on estime actuellement à 70 % le taux global de remboursement par la Sécurité sociale (protection de base).

Graphique 1. La couverture maladie obligatoire en 1960 et en 2000 (effectif en %)



Graphique 2. Financement des dépenses médicales en 1960 et 2000 (en %)



L'évolution en 40 ans est considérable, puisqu'une personne sur quatre n'était pas protégée par un régime obligatoire en 1960 (dont 14 % n'avait aucune couverture maladie), alors que tout le monde l'est en 2000 ; le régime des indépendants est créé ; les régimes spéciaux deviennent marginaux ; le régime général couvre 84 % de la population (Cf. gr. 1).

A cette protection de base s'ajoute une protection maladie complémentaire (mutuelle ou assurance privée) qui rembourse tout ou partie des dépenses médicales non prises en charge par la Sécurité sociale.

En matière de financement aussi, l'évolution est importante avec l'augmentation de la part de la Sécurité sociale (tous régimes confondus) et des couvertures complémentaires, le désengagement apparent de l'Etat, la diminution de la parts des ménages.

Certaines mutuelles de fonctionnaires, qui existaient avant la guerre, ont obtenu, lors de la création de la Sécurité sociale, de continuer à gérer entièrement les prestations de leurs adhérents (partie obligatoire et complémentaire). Ces mutuelles dites « mutuelles décompteuses » ont un mode de gestion distinct de celui des CNAMTS, avec des systèmes informatiques différents.

Les fichiers de Sécurité sociale contiennent les renseignements nécessaires au versement des prestations :

- en ce qui concerne les personnes : identité, domicile et/ou lieu de versement des prestations, âge et sexe, éléments justifiant les droits ouverts (nature de la protection, éléments de modification tels que maternité, invalidité, ou affections de longue durée),
- en ce qui concerne les prestations : nature et lieu de la consommation, nature du producteur.

2 *Décision du ministre Jean-Noël Jeanneney supprimant la FNOSS et la remplaçant par trois caisses : la CNAMTS (maladie), la CNAVTS (vieillesse) et la CNAF (famille).*

2. Les premiers sondages

1.1. Une enquête sur 400 000 cas de maladie (1952-1953-1954)

Il semble que ce soit véritablement le tout premier sondage dans les dossiers de Sécurité sociale depuis sa création en 1945. Il s'agit d'un sondage au 1/100ème effectué sur les dossiers des caisses primaires d'assurance maladie (CPAM) de Sécurité sociale. 68 caisses primaires (sur 122) ont participé au sondage (129 505 personnes, 393 256 maladies), dont les plus importantes (Région parisienne). Les dossiers ont été vus par les médecins contrôleurs de la FNOSS et suivis pendant 3 ans, 1952-1953-1954, toutes les prestations ont été relevées et affectées aux maladies.

Statistiques publiées : ventilation Paris - province ; nombre de maladies selon l'âge et le sexe, nombre d'épisodes morbides, montant des prestations de Sécurité sociale par personne selon l'âge et le sexe, selon le nombre de maladies et selon leur nature par grands chapitres [1].

1.2. L'enquête dans les dossiers médicaux de la SNCF (1956)

Il s'agit d'un sondage effectué sur les dossiers de la Caisse de prévoyance de la SNCF[3] (qui joue pour les cheminots le rôle de Sécurité sociale). Sondage au 1/320, soit 1792 dossiers, examinés un par un par un médecin, qui remplissait des fiches individuelles. Relevé des données individuelles connues de la Caisse de prévoyance, de la morbidité, et des consommations médicales du 3ème trimestre de 1956 [12].

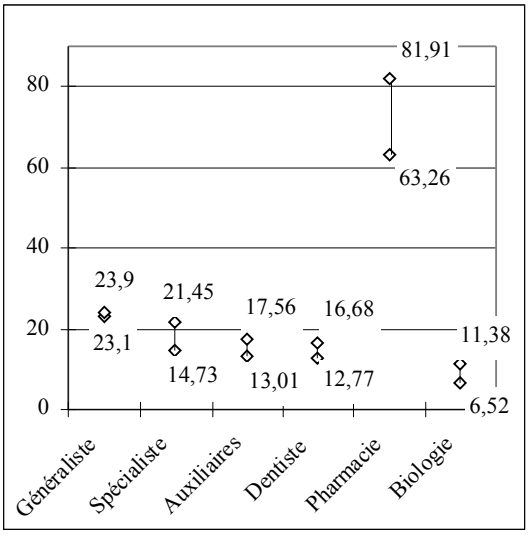
1.3. L'enquête pilote de 1965

Le premier appariement en France entre données d'enquête et données de Sécurité sociale a été effectué en 1965 sur une petite enquête expérimentale [13]. C'était une observation des consommations médicales, des maladies et du contexte socio-économique sur 549 ménages de la Région Parisienne (377 participations complètes, 67 abandons, 67 refus, 38 introuvables). L'objectif était de tester la possibilité et de mettre au point les méthodes et les documents (questionnaires, cartes réponses…) d'une enquête de plusieurs mois, avec plusieurs visites d'enquêteur. Les N° de Sécurité sociale ont été demandés aux enquêtés au cours de l'enquête, ils ont été fournis sans problème ; un attaché de l'INSEE a relevé, sur la même période tous les remboursements dans les dossiers de Sécurité sociale ; il les a ensuite appariés manuellement **individu par individu** et **acte par acte** avec les consommations médicales relevées au cours de l'enquête ménages ; il pour cela disposait de données très fines relevées dans l'enquête (lieu, date, nature du producteur, contenu de l'acte, motif, prix payé,…). Un travail aussi fin et détaillé n'a jamais pu être refait. Notons que la loi sur les données personnelles n'existait pas et que la protection des personnes résultait de la déontologie habituelle.

La confrontation des deux sources de données a fourni des estimations sur les taux d'erreurs. Ces biais ont deux origines : la plus grande consommation médicale des non-répondants, les oublis, erreurs et fausses déclarations [6].

3 *Société national des chemins de fer*

Lien entre niveau de soins médicaux et participation à l'enquête



Les personnes qui refusent de participer à l'enquête ou abandonnent ont une consommation de soins de ville (d'après les dossiers de Sécurité sociale) globalement supérieure aux participants à l'enquête de 30 %.
Cette « surconsommation » dépend des soins considérés, de 3 % pour les soins de généraliste, 30 % pour la pharmacie prescrite, 46 % pour les soins de spécialistes, elle atteint 74 % pour les analyses de laboratoire[4]. Les non participants sont plus vieux que les participants, et leur morbidité est supérieure ; si on corrige de cet effet de l'âge, la « surconsommation » des refus et abandons est réduite, mais encore supérieure de 21 % aux participants. Au total, corriger les données d'enquête de la structure de l'échantillon en termes d'âge et de sexe (redressement), en la rapportant à celle de la population générale ne corrige qu'une partie du biais ; ce problème est doublement grave, car l'erreur n'est pas seulement en niveau mais aussi en structure

Oublis et négligences : certains enquêtés ne déclarent pas toutes leurs consommations. L'estimation des taux de sous déclaration est complexe et dépend de la qualité de l'appariement entre données d'enquête et données de Sécurité sociale. Quatre cas sont possibles :

- les actes appariés,
- les actes déclarés dans l'enquête, non retrouvés dans les dossiers de Sécurité sociale,
- les actes dans les dossiers mais non déclarés dans l'enquête,
- les inconnus complets.

Un acte mal relevé peut ne pas être apparié alors qu'il a bien été déclaré par les enquêtés (erreur dans le relevé d'enquête ou dans le dossier de Sécurité sociale, sur la date, le lieu, la nature,…) et réciproquement, un acte peut avoir été apparié à tort ; du fait de la sévérité de nos critères d'appariement, nous avons considéré comme négligeable les faux appariements[5], ce qui nous a permis d'estimer une fourchette des taux d'oublis (globalement, entre 14 % et 38 % des actes intégralement payés déclarés) ; réciproquement, on pouvait estimer le taux d'actes ne figurant pas dans les dossiers de Sécurité sociale entre 1 % et 22 %. Les taux de sous estimation dépendent de la nature de la consommation (pouvant atteindre, dans les données d'enquête, 28 % pour les soins de spécialiste et 50 % pour ceux d'auxiliaire, et dans les données de Sécurité sociale, 17 % pour la biologie et 25 % pour les soins de spécialiste) : comme pour les biais dû aux refus, l'erreur est double, en niveau et en structure [6].

Cette enquête pilote a permis de mettre au point les futures enquêtes auprès des ménages sur la santé et les soins médicaux, dont la méthodologie a été conservée en 1970, 1980 et 1991, elle a été modifiée en 2002. Elle a aussi fourni des ordres de grandeur des biais sur la consommation médicale estimée à partir des enquêtes auprès des ménages (utilisés en comptabilité nationale). Il faut noter que le succès de cette opération pilote repose en partie sur l'adhésion des enquêteurs de l'INSEE, triés sur le volet et très enthousiastes à l'idée de participer à une enquête expérimentale de cette importance ; ce succès repose aussi sur le fait que l'appariement a été fait manuellement par un attaché de l'INSEE, détaché dans les locaux de la Sécurité sociale dont l'aide lui a été précieuse. Nous n'avons pas su renouveler cet appariement, au niveau le plus élémentaire.

Ce non renouvellement est tout à fait regrettable, car si l'enquête auprès des ménages est sous estimée, les données de Sécurité sociale ne sont pas non plus exhaustives :

- certains actes, peu remboursés, ne sont pas présentés au remboursement ; de plus, les prix de ces consommations non connues de la Sécurité sociale sont différents de ceux des soins remboursés ; il s'ensuit une sous estimation de la consommation médicale et, selon les cas, surestimation des prix
- certains actes donnent lieu à des dépassements non déclarés à la Sécurité sociale entraînant une sous estimation des prix et des dépenses des malades.

Ces biais d'estimation sont rendus plus gênants du fait qu'ils ne sont uniformes ni selon les types de soins, ni selon les catégories de patients (morbidité, âge, sexe, niveau social, revenu,…) ; ces biais ne sont pas non plus

4 *rappelons que les données datent de 1965 et que la structure de la consommation médicale n'était pas la même qu'aujourd'hui, non plus que les taux de remboursement*

5 *actes différents relevés dans l'enquête ménage et dans les dossiers de Sécurité sociale, considérés comme un même acte*

stables dans le temps, ce qui rend les études d'évolution fragiles, surtout lorsqu'on s'intéresse à des périodes relativement longues (plusieurs années).

C'est pourquoi il serait utile de refaire ces appariements au niveau élémentaire à intervalles réguliers, en dépit du coût et de la difficulté des procédures. zzz

1.4. Les enquêtes de 1966-67 au 12ème et au 60ème

Deux sondages gigognes ont été ont été effectués dans les données de Sécurité sociale (quasi-totalité des régimes) en 1966-67 [2].Les informations étaient relevées dans une optique comptable : des compteurs étaient définis dans un « plan statistique », correspondant aux variables dont on voulait estimer la dépense (par exemple, consultation de généraliste pour un homme >=60 ans, …) ; chaque prestation versée à un assuré appartenant à l'échantillon venait incrémenter les compteurs correspondants (actes médicaux de la nomenclature des actes professionnels) au niveau des centres de paiement. Ces sommations étaient regroupées dans chacune des 122 caisses primaires pour créer des tableaux envoyés à la FNOSS, qui les totalisait au niveau national et les regroupait avec les données des autres régimes pour établir des statistiques mensuelles et annuelles[6]. Les tableaux étaient moins détaillés dans le sondage au 12ème (*assurés nés au mois d'octobre, et leurs ayants droit*) que dans le sondage au 60ème (*assurés nés au mois d'octobre d'une année se terminant par 4 ou 9, et leurs ayants droit*).

1.5. L'enquête nationale de 1970

L'enquête pilote de 1965 a permis de décider les modalités de l'enquête nationale de sur la santé et les soins médicaux 1970 (qui ont été reprises en 1980 et 1991) : durée d'enquête prolongée à trois mois, nombre et durée des interviews (5 entretiens d'environ 3/4 d'heures), nature et formalisation des questionnaires (« initial » – *orienté vers les variables socioéconomiques* – « ramassage » – *orienté vers les soins médicaux* – « navette » – *orienté vers les évènements dans la durée* – « final » – *variables personnelles* -) et des autres documents [7]. Les résultats d'appariement obtenus lors de l'enquête pilote ont incité à rééditer la procédure d'appariement, mais cette fois au plan national (23197 personnes) [7].

Pour aider à réaliser l'appariement, les NNI (numéros nationaux d'identité[7]) sont relevés auprès des enquêtés et recherchés dans les caisses primaires du seul régime général (des travailleurs salariés). Toujours pas de CNIL, ni de consigne particulière pour cette opération, qui pourtant, plus complexe, puisque exécutée au plan national et de manière décentralisée, présentait un plus grand risque de dérapage.

Etant donnée la grande taille des fichiers, et les progrès réalisés en informatique, il a été décidé que l'appariement serait fait de manière centralisée et en informatique : tous les dossiers de versements concernant les ressortissants du régime général[8] Sécurité sociale (des travailleurs salariés) enquêtés et retrouvés ont été photocopiés dans les centres de paiement de la Sécurité sociale et envoyées à l'INSEE pour appariement. Ces documents étant très hétérogènes (entre Caisses primaires d'assurance maladie), un quasi questionnaire a été créé et renseigné à l'INSEE. L'appariement devait être fait entre ces quasi questionnaires et les questionnaires d'enquête.

L'appariement a été un échec, non seulement au niveau élémentaire des consommations et des remboursements pour lesquels aucun appariement n'a pu être fait, mais même au niveau des personnes, puisque sur 23 197 personnes enquêtées sur le terrain, près de deux tiers seulement étaient appariables (14 733 personnes, soit 64 %), un tiers seulement a été apparié (7 393 personnes, soit 32 %). Ce travail a été considéré comme un échec et non publié ; les difficultés étaient considérables et nous n'avons-nous pas su les surmonter ; peut être aurait-il mieux valu faire ce travail manuellement, et de manière décentralisée dans les Caisses Primaires de Sécurité sociale, comme cela avait été fait dans l'enquête pilote de 1965.

Pourrait-on aujourd'hui faire un appariement informatique ? Depuis 1970, un important effort d'homogénéisation des procédures de Sécurité sociale (documents administratifs, nomenclatures,…) a été fait à l'instigation des statisticiens, et les moyens informatiques actuels sont incommensurables par rapport aux moyens de calculs disponibles à cette époque (cartes perforées, trieuses, tabulatrices).

---

6 *Ce travail, a été initié par Antoine Sanson-Carette, administrateur INSEE, détaché au Ministère des Affaires sociales et René Nathan statisticien à la FNOSS (fédération nationale des organismes de sécurité sociale), et était réalisé sous l'égide du « Groupe d'harmonisation des statistiques de Sécurité sociale (Régime général, Régime agricole, Régime des travailleurs non salariés, Régime militaire, Régime de la SNCF, Régime de la RATP, Régime de la marine marchande et Régime des mines).*

7 *Voir plus loin, § 2.3..*

8 *Et quelques régimes rattachés au Régime général (fonctionnaires, EDF-GDF, étudiants, invalides de guerre)*

Il faut souligner qu'à ce moment, la base d'observation était l'enquête auprès des ménages, l'apport des données de Sécurité sociale ayant essentiellement pour but d'en mesurer l'erreur, éventuellement de les compléter.

L'échec de l'appariement de 1970 provenait en partie du fait que les personnes enquêtées étaient celles appartenant à des ménages tirés au sort à partir du recensement et dont on ne retrouvait pas trace dans les fichiers de Sécurité sociale : au biais découlant des refus de participation venait s'ajouter celui de l'absence de dossier.

2. L'échantillon permanent d'assurés sociaux

2.1. La situation statistique dans les années 1970

Les données sur lesquelles nous travaillions provenaient des enquêtes décennales auprès des ménages (1960, 1970, 1980,…) ; l'évolution du secteur médical était trop rapide pour pouvoir la suivre convenablement avec des intervalles de 10 ans.

Deux éléments ont permis d'avancer : l'idée que les données de Sécurité sociales, même non reliées aux données d'enquête auprès des ménages pourraient être utiles aux recherches et l'intention d'abandonner le sondage à partir du recensement pour le faire directement sur les données de Sécurité sociale.

Les données comptables de Sécurité sociale, exhaustives et peu maniables sont très lourdes à mettre en place, et les responsables du Régime Général ont donc accueilli favorablement, en 1975, le projet d'un sondage conduisant à un échantillon de taille réduite permettant d'obtenir des estimations au fur et à mesure des besoins.

2.2. Les objectifs, les obstacles

Les objectifs de ce sondage étaient multiples et différents selon les acteurs :
- créer plus rapidement, en routine ou au fur et à mesure des besoins, les statistiques nécessaires à la gestion courante de la Sécurité sociale,
- établir des simulations pour évaluer l'impact des nouveaux projets de la de Sécurité sociale, en particulier lors des négociations avec les syndicats des professionnels de santé,
- fournir une base statistique pour la recherche économique ; dans ce but, il a été prévu de faire une enquête complémentaire auprès des ménages pour relever les informations nécessaires inconnues de la Sécurité sociale,
- étudier l'évolution de la consommation médicale d'une même personne au cours du temps (panel).

Les contraintes et exigences découlant de l'organisation des fichiers de Sécurité sociale et des moyens disponibles (possibilités financières, moyens informatiques, personnels,…) ont conduit aux choix suivants :
- l'échantillon est tiré dans les fichiers de Sécurité sociale et non du recensement (comme dans l'enquête nationale de 1970), il ne peut donc s'agir que d'un **sondage en grappe** (l'assuré et ses ayants droit),
- l'échantillon est compris entre 50 et 60 000 personnes**,**
- la méthode de sondage doit permettre de relever tous les remboursements pour les assurés appartenant à l'échantillon et regrouper les remboursements d'un même assuré quelque soit l'organisme payeur,
- l'échantillon est mis à jour régulièrement (départ des sortants, introduction des entrants) : il évolue comme la population de référence.

L'origine principale des difficultés de ce projetrésidait dans la diversité des institutions de la Sécurité sociale française[9], dans leur autonomie de gestion et de fonctionnement et corrélativement dans la multiplicité et la dispersion des fichiers d'immatriculation des assurés sociaux et des prestations (les mêmes prestations donnaient lieu à des enregistrements voire des nomenclatures différents) [9].

Les documents concernant la même personne peuvent se retrouver dans différents centres de paiement si elle a déménagé et que la régularisation n'est pas achevée.

En deuxième lieu, une même personne peut avoir des droits différents si, travaillant (simultanément ou à intervalles rapprochés) sous deux statuts différents (par exemple salarié agricole l'été et salarié du tourisme l'hiver) elle est assujettie à des régimes de Sécurité sociale différents.

Par ailleurs, une personne peut être ayant droit de l'un de ses parents (*ou de son conjoint*) et en même temps assurée à titre personnel ; elle a alors un dossier sous l'identifiant du parent (*ou du conjoint*) dont elle est l'ayant droit et un deuxième dossier sous son propre identifiant. Elle apparaît donc dans deux dossiers d'assurés, sous deux identifiants différents, qu'on ne peut regrouper. Or on a besoin d'identifier de manière univoque chaque personne protégée et de lui attribuer toutes les prestations dont elle a bénéficié, et en même temps d'estimer correctement le nombre de personnes protégées.

La troisième difficulté résidait dans le fait qu'en 1977 la « liquidation » des dossiers (contrôle, calcul et versement des prestations) était effectué manuellement, comme d'ailleurs la gestion de l'ensemble des fichiers (immatriculation, modifications de droits,…).

---

9 *Cf. §. 1*

La méthode de sondage devait permettre de relever manuellement tous les remboursements à des assurés appartenant à l'échantillon et regrouper les remboursements d'un même assuré quelqu'en soit l'organisme payeur ; nous espérions aussi que notre échantillon contiendrait des non prestataires au même titre que des prestataires.

2.3. La méthode de sondage

Le principe de la méthode consiste à définir une condition restrictive C, pouvant être mise en œuvre dans tous les fichiers ; l'échantillon est constitué des les personnes vérifiant la condition C. La proportion, dans la population, de personnes vérifiant la condition C égale le taux de sondage [11].

Pour être utilisable, la condition **C** doit avoir les quatre propriétés suivantes :

a. la condition **C** doit s'appliquer à un caractère se trouvant dans tous les fichiers et devait être facilement accessible manuellement[10],

b. les personnes vérifiant **C** doivent être uniformément réparties dans la population, de manière que l'échantillon soit un sous ensemble représentatif, en particulier la répartition doit être uniforme géographiquement,

b. la condition **C** doit s'appliquer à un caractère stable dans le temps pour que toutes les générations soient également représentées et, avec une mise à jour automatique, que la représentativité continue d'être assurée,

d. **C** doit être stable dans l'espace (des fichiers), autrement dit apparaître de la même manière dans les différents fichiers ; si une personne vérifiant C dans un fichier figure dans un deuxième fichier, il faut qu'elle appartienne aux deux sous échantillons issus de ces deux fichiers : par regroupement (sous le même identificateur) et apurement, on est ainsi en mesure d'estimer la consommation totale de chaque personne quelques soient les fichiers dans lesquels elle apparaît.

L'intérêt de ce mode de sondage est qu'il conduit à un échantillon en permanence représentatif de la population de base ; la proportion de personnes vérifiant **C** est stable au cours du temps, ainsi que parmi les départs (décès) et parmi les entrants (naissances[11]). L'échantillon évolue comme la population de référence.

Les seules informations existant toujours dans tous les fichiers sont le (ou les) nom et le (ou les) prénom. Nous pensions pouvoir utiliser cette particularité pour faire un sondage à partir de l'un ou de l'autre ; par exemple, choisir l'initiale du nom (ou un groupe de lettres) issue (s) du nom. Toutes les lettres ne sont pas aussi fréquentes et nous envisagions de choisir une lettre, ou un groupe de lettre, dont la fréquence correspondrait au taux de sondage visé.

L'INSEE nous a fourni un fichier de 103 300 enregistrements représentatif des personnes nées en France, tiré au hasard dans la base 1975 de SAFARI (système automatisé pour les fichiers administratifs et le répertoire des individus), qui recense toutes les personnes résidant en France, relève leurs lieu et date de naissance, de mariage, de décès… leur affecte le NNI[12]. La CNAVTS nous a fourni un fichier de 48 119 personnes nées hors de France, représentatif de son fichier national. La validation du sondage a été effectuée sur ces deux fichiers : les noms ne sont pas uniformément répandus sur le territoire (origines ethniques) car ils sont distribués différemment selon les régions de France, quant aux prénoms, ils sont soumis à des effets de mode et on ne peut retenir l'hypothèse d'une stabilité dans le temps (Cf. tableau 1). Une tentative sur la deuxième lettre donne des résultats comparables ;: certains noms ne comportent pas plus de deux lettres, on ne peut pas aller plus loin. Les propriétés d (stabilité dans l'espace) et b (stabilité dans le temps) de la condition **C** ne sont pas respectées, et ni les noms, ni les prénoms ne peuvent être utilisés pour un sondage représentatif.

---

10 *Comme on envisageait déjà l'informatisation à venir, **C** devait pouvoir aussi être traitée de manière informatique*

11 *Cette propriété est vraie hors migration :si les entrants ont intérêt à figurer dans les fichiers d'immatriculation,et le font dès leur premier problème de santé, les partants n'ont aucune raison de le signaler à leur caisse de Sécurité sociale*

12 *Voir plus loin, § 2.3..*

Tableau 1 : Initiales du nom et du prénom les plus fréquentes dans les régions et les générations où elles sont minimum et maximum

	Initiale du nom					Initiale du prénom			
Région	**B**	**L**	**C**	**D**	**décennie de naissance**	**M**	**J**	**A**	**C**
Est	**10,6**	6	6,5	5,7	**1890/1899**	**25,5**	11,9	12,2	4,8
Ouest	13,2	**14,3**	**8,7**	6,5	**1910/1919**	19,8	11,8	**12,7**	**3,8**
Centre est	**13,8**	**5,2**	8,1	**8,1**	**1940/1949**	18,9	**18,5**	9,6	8,6
Région méditerranéenne	12	**5,2**	**5,6**	**5,6**	**1960/1969**	10,3	8,3	**5,7**	11,5
Ensemble	*12,3*	*10*	*9,3*	*5,6*	**1970/1977**	**7,9**	**5,6**	6	**12,1**
					Ensemble	*16,4*	*12,8*	*8,9*	*7,9*

Remarque anecdotique : le prénom le plus fréquent était cette année Jean pour les hommes et Marie pour les femmes, le nom le plus répandu était Durant[13] ; ce résultat sur les noms a été repris et étendu dans des publications sur la fréquence des noms, et sur leur évolution (noms qui se développent ou qui disparaissent), et a peut être influencé la nouvelle législation sur la transmission des noms ; quant au résultat sur les prénom, il est régulièrement repris dans des publications sur les prénoms à la mode, pour aider les futurs parents à choisir le prénom de leur enfant.

L'échec de l'utilisation du nom ou du prénom a été une grande déception, car il n'existait aucune autre information commune sur les personnes protégées dans les fichiers de tous les régimes de Sécurité sociale.

Nous avons dû nous rabattre sur un deuxième identifiant utilisé à l'époque par le seul Régime général, le Numéro national d'identité (NNI). Les ambitions du sondage ont donc été réduites à ce seul régime. Nous espérions qu'à terme, tous les régimes utiliseraient cet identificateur.

Depuis cette période, les deux autres grands régimes (agricoles et indépendants) ont introduit le NNI dans leurs fichiers et se sont joints au sondage actuel ; les trois régimes représentent actuellement environ 95 % de la population résidant en France.

Le NNI, de 15 caractères, est constitué de 6 zones :

sexe	deux derniers chiffres de l'année de naissance	mois de naissance	département (ou pays) de naissance	commune de naissance	numéro d'ordre	clé de contrôle
□	□□	□□	□□	□□□	□□□	□□

Le sexe prend les valeurs 1 (hommes) et 2 (femmes). Les deux derniers chiffre de l'année de naissance posent problème pour les centenaires qui sont de plus en plus nombreux (problème néanmoins encore marginal). Le mois de naissance prend les valeurs de 1 à 12, sauf pour certains immigrés (3799, soit 7,7 %) dont on ne connaît pas le mois de naissance ; on trouve ainsi dans le fichier des NNI des immigrés 16 valeurs supérieures à 12 comme mois de naissance, sans que nous ayons pu en déterminer l'origine, qui ne semble pas lié au pays de naissance. Les départements ont des numéros à deux chiffres (sauf la Corse ventilée en deux départements, 28A et 28B) et dans chaque département, les communes sont classées de 1 à n, en général par ordre alphabétique. Le numéro d'ordre est donné par l'INSEE à la naissance, les enfants étant classés par commune, par jour de naissance et par ordre alphabétique. La clé de contrôle a été introduite pour vérifier de manière informatique l'ensemble du NNI. En 1977, la clé n'apparaissait pas sur tous les documents, comme elle le fait depuis l'informatisation complète du système.

13 *LEVY, M. L.,* Répartition du mois de naissance et des initiales en France. *Population, mai – juin 1980, pages 708-711*

CONDITION DE COMPATIBILITÉ

Pour enrichir les informations, et plutôt que de faire une enquête ad hoc sur les personnes protégées du sondage, nous espérions nous rattacher à l'enquête «salariés 3» de l'INSEE, pour être compatibles avec cette enquête, il fallait prendre pour année de naissance les années paires et pour mois de naissance le mois d'octobre. Ces deux contraintes représentaient la « condition de compatibilité » (avec l'enquête INSEE). La structure des mois de naissance est relativement stable sur le territoire français ; au cours du vingtième siècle, la saisonnalité des naissances a légèrement varié et le mois de naissance peut entraîner un léger biais sur la représentation des âges, dont l'effet est négligeable. Pour la représentativité des personnes nées hors de France, nous avons ajouté à la valeur 10 deux valeurs de mois supérieures à 12 (30 et 50) ; On obtient une bonne représentativité selon le sexe, une légère sur représentation des personnes les plus âgées, et des personnes nées en Algérie et dans les DOM-TOM. Les naissances au mois d'octobre sont un peu moins fréquentes que la moyenne des autres mois, se situant, selon la décennie, entre 7,83 % et 8 % des naissances annuelles au cours des dernières décennies (alors que la proportion moyenne, pour 31 jours, serait de 8,49 %). Les années paires ne posent aucun problème de représentativité ; en revanche, au moment de l'exploitation, on ne peut pas faire des classes de 5 ans de tailles égales.

Cette double contrainte conduit, avec un taux de sondage d'environ 1/24, à un échantillon beaucoup trop important pour que nous puissions le traiter avec nos moyen informatiques de 1980.

Il fallait donc introduire de nouvelles contraintes. On ne peut sonder sur le sexe car on désire avoir des informations sur les deux sexes. On ne peut pas sonder sur le département (ou le pays) de naissance, car on veut tous les avoir. On ne peut pas sonder sur la commune de naissance, car on risquerait de perdre certaines communes importantes (par exemple une métropole régionale). Reste le numéro d'ordre : le nombre de naissances est très dispersé selon la taille de la commune (dans les communes à forte densité de population, les naissances sont nombreuses, et les numéros d'ordre peuvent atteindre plusieurs centaines par jour ; au contraire, dans les communes rurales, on a un petit nombre de naissances, voire pas du tout) : choisir un numéro d'ordre petit donnerait un poids trop grand aux communes rurales, un numéro d'ordre grand les ferait disparaître complètement. La clé de contrôle est aujourd'hui partie intégrante du NNI ; elle ne l'était pas en 1977 (elle était en cours d'introduction en vue de l'informatisation prévue).

LA CONDITION SPÉCIFIQUE

Nous nous sommes donc résignés à définir une méthode de sondage un peu plus compliquée [11] : nous avons déterminé une condition satisfaisant aux conditions a, b, c, d et d'un emploi suffisamment simple pour être opérationnelle dans la sélection manuelle (et ultérieurement informatique) des assurés de l'échantillon et de tous les documents les concernant : les deux derniers chiffres du numéro de commune de naissance sont égaux aux deux derniers chiffres du numéro d'ordre. En moyenne et pour chaque mois, parmi les petites communes, avec peu de naissances, seules sont représentées celles se terminant par 01, éventuellement par 02 s'il y a 2 naissances ; parmi les communes ayant n naissances, les communes se terminant par 01 sont représentées par la première naissance, celles se terminant par 02, par leur deuxième naissance,… celles se terminant par n par leur $n^{ème}$ naissance ; les communes ayant plusieurs centaines de naissances sont représentées plusieurs fois (autant que de centaines). Cette condition conduit à un taux de sondage de 1/100, qui, composé avec la condition de compatibilité conduit à 1/2400 environ, trop éloigné du taux recherché (environ 1/1000, soit un échantillon d'environ 60000 personnes. Pour redoubler l'échantillon nous avons introduit la condition inverse : les deux derniers chiffres du numéro de commune sont égaux aux deux derniers chiffres du numéro d'ordre, dans l'ordre inverse (par exemple 37 et 73). La condition spécifique, réunion de ces deux conditions, deux derniers chiffres du numéro de commune de naissance égaux aux deux derniers chiffres du numéro d'ordre, dans l'ordre ou dans le désordre, conduit à un taux de sondage de 1/50, qui composé avec la condition de compatibilité conduit à 1/1200 environ, taux retenu.

2.4. Validation

Au total, la condition **C**, réunion de la condition de compatibilité (assurés nés au mois d'octobre d'une année paire) et de la condition spécifique (deux derniers chiffres du numéro de commune sont égaux aux deux derniers chiffres du numéro d'ordre dans l'ordre inverse) conduisait en 1975,

- pour la population née en France, à un échantillon satisfaisant aux conditions présentées plus haut : universalité (le NNI apparaît sur tous les bordereaux de prestation), stabilité dans le temps (taux de sondage est indépendant de la classe d'âge) et dans l'espace (le taux de sondage est indépendant de la région), représentativité (toutes les catégories de communes sont également représentées) ; pour toutes ces conditions, la probabilité d'erreur est inférieure à 5 1/1244,
- pour les personnes nées hors de France, à une bonne représentativité selon le sexe, une légère sur représentation des personnes les plus âgées et des personnes nées en Algérie et dans les DOM-TOM ; le taux de sondage était de 1/1052.

Le taux de sondage global était de 1/1222 et à un échantillon d'environ 60 000 000 / 1200 = 50 000 personnes proches du taux visé (1/1200), avec une légère sur représentation des personnes nées à l'étranger [11]. Nous pensions que ce biais devait se réduire au cours du temps, avec le développement des états civils dans les pays en voie de développement, mais nous n'avons pas contrôlé à nouveau cette hypothèse.

Nous avons vérifié dans un centre de paiement que la reconnaissance des pièces à retenir est facile à faire lors du règlement des dossiers et que l'indication du NNI, obligatoire pour obtenir les remboursements est bien dans tous les fichiers [13].

LES PROBLÈMES NON RÉSOLUS :

- erreurs à l'entrée : pour pouvoir être soignées, certaines personnes, non encore immatriculées à la Sécurité sociale, reçoivent un numéro provisoire : pour l'année en cours, leurs documents se partagent entre numéros provisoire et définitif, ce cas est relativement marginal,
- erreurs à la sortie : les personnes décédées ne sont retirées des fichiers qu'avec beaucoup de retard, et les personnes parties à l'étranger ne le sont souvent pas du tout,
- perte de consommation par immatriculations fictives et « forçage » : lorsque certaines erreurs sur le numéro d'immatriculation (NNI) ne permettent pas de retrouver la personne, les liquidateurs sont autorisés à « forcer » le versement des prestations en faisant appel à des numéros fictifs ; ce cas n'est pas très fréquent,
- protections multiples ne pouvant pas être regroupées (surestimation du nombre de personnes protégées, sous-estimation de la consommation :
    - personne protégée par deux régimes dont un non participant au sondage,
    - enfant pouvant être ayant droit de leurs deux parents : du fait des origines historiques et professionnelles de la Sécurité sociale, seuls les assurés (actifs, retraités, chômeurs,…) sont directement protégés, les autres (enfants, conjoints inactifs) sont leurs ayants droit ; si ses deux parents sont actifs, un enfant peut être ayant droit de l'un, de l'autre ou des deux ; dans ce dernier cas, les prestations relatives à sa consommation sont ventilées dans les dossiers de ses deux parents[14],
    - Un assuré peut être en même temps ayant droit : lorsqu'un enfant devient actif, la Sécurité sociale lui ouvre un dossier à titre personnel, il devrait être supprimé du dossier de son parent dont il était ayant droit, ce qui n'est pas toujours le cas ; il en est de même d'une personne ayant droit de son conjoint devenant active

      Ces problèmes devraient être résolus avec la prochaine réforme, à l'issue de laquelle chacun devrait être personnellement assuré et remboursé sous son propre NNI.

2.5. La mise en place et l'extension de l'EPAS (échantillon permanent d'assurés sociaux)

Le sondage (au 1/1200) a commencé de fonctionner le 1er janvier 1978, sur sept CPAM[15], de Normandie et un département de la même région pour les Régimes des agricoles et des indépendants. Dans tous les centres de paiement, la saisie était manuelle.

Lorsque, trois ans plus tard, le recueil a été informatisé, les systèmes informatiques des Régimes des agricoles et des indépendants étant différents, ces deux régimes se sont retirés de l'opération. Au niveau du Régime général cependant, le champ géographique s'est progressivement étendu à 46 CPAM au 1er janvier 1981, 106 CPAM au 1er janvier 1985 et à l'ensemble des CPAM (122) au 1er janvier 1989 : 11 ans de montée en charge pour le seul Régime général ! La difficulté était technique (certaines CPAM avaient des systèmes informatiques différents du système national) et politique (convaincre les responsables des CPAM qu'ils ne perdaient pas une trop grande part d'autonomie[16].

L'EPAS ne couvre pas l'ensemble du Régime des salariés, puisque les mutuelles de fonctionnaires[17] (« mutuelles décompteuses »), qui gèrent l'ensemble des remboursements de leurs membres, parties obligatoire et complémentaire, ne désirent pas s'associer à ce sondage

---

14 *Il peut se produire que les prestations soient toujours demandées sous le NNI d'un seul des deux parents, l'enfant n'en demeure pas moins ayant droit de l'autre parent, il est alors considéré à tort comme non consommateur*

15 *Les CPAM, caisses primaires d'assurance maladie, sont les organismes de base du Régime général des travailleurs salariés, en général une par département, elles sont autonomes (et dotées d'un Conseil d'Administration)*

16 *cette tâche difficile a été effectuée par René Nathan chef du service statistique de la CNAMTS et Alain Gaillard, statisticien*

17 *Cf. partie 1*

En sens inverse, le Régime des indépendants (CANAM) s'est associé au sondage en 1994 et le Régime des agricoles (MSA) en 1996. Aujourd'hui, le sondage EPAS couvre environ 95 % de la population, en sont exclus les adhérents des « mutuelles décompteuses[18] » et les assurés des régimes spéciaux (Régime militaire, Régime de la SNCF, Régime de la RATP, Régime de la marine marchande et Régime des mines).

L'EPAS a été doublé en 2000, le taux de sondage passant à 1/600, en prenant toutes les années de naissance (en supprimant la condition, naissance une année paire) soit un échantillon d'environ100 000 personnes.

3. L'enquête auprès des ménages, (ESPS, Enquête santé et protection sociale)

Les fichiers de Sécurité sociale contiennent les renseignements nécessaires au versement des prestations :

- en ce qui concerne les personnes : identité, domicile et/ou lieu de versement des prestations, âge et sexe, droits aux remboursements, …
- en ce qui concerne les prestations : nature et lieu de la consommation, nature du producteur, tarif, taux de remboursement, prix payé,….

Si on veut utiliser ces informations à d'autres fins[19] que le versement des prestations, par exemple pour la recherche, il faut les transformer, éventuellement leur adjoindre d'autres variables utiles à ces fins (variables relatives au patient ou au producteur).

Dès la création de l'EPAS, nous avions envisagé d'adjoindre aux données du sondage sur les dossiers, des informations issues de l'enquête emploi de l'INSEE, mais, après quelques promesses et de longues hésitations, nous n'avions pas pu obtenir l'accord des responsables de cette enquête ; L'INSEE était en effet hostile à ce moment à la fusion de fichiers pour des raisons tant déontologiques que pratiques.

Nous avons donc prévu une enquête ad hoc auprès des assurés de l'EPAS. Dès 1985, à partir du moment où la grande majorité des CPAM (106 sur 122) ont effectivement participé au sondage dans les dossiers d'assurés, le projet d'enquête complémentaire ménages est devenu consistant. Une enquête expérimentale a été effectuée en 1986 ; plusieurs points devaient être élucidés : nature du réseau d'enquêteurs (agents de la CNAMTS ou enquêteurs professionnels), durée de l'observation (15 jours ou un mois), mode de contact (téléphone ou face à face), champs des informations à recueillir (socio-démographie, morbidité, protection maladie complémentaire, consommation médicale déclarée) et formulation des questionnaires [13].

Pour tenir compte des problèmes de confidentialité une procédure d'appariement des individus en double aveugle a obtenu l'accord de la CNIL.

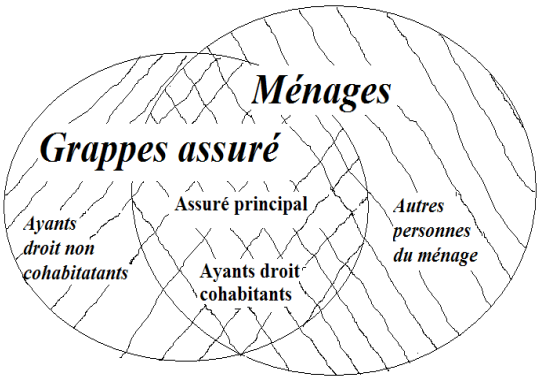
L'enquête auprès des ménages (ESPS, enquête sur la santé et la protection sociale) a démarré le 1er janvier 1988 [3]. Les budgets ne permettaient pas d'enquêter 50 000 personnes tous les ans, aussi porte-t-elle sur un quart de l'EPAS chaque année (20 à 25 000 personnes). Cette enquête, par téléphone et courrier ou en face à face (pour les personnes non contactées par téléphone), portait au départ sur les variables socio-démographiques, la protection maladie non obligatoire, la morbidité et la consommation médicale du mois précédent ; le questionnaire comprenait des questions d'opinion et un module libre était prévu pour d'éventuelles questions qui se poseraient chacune des années suivantes. Depuis, le questionnaire a un peu évolué, mais permet des analyses longitudinales.

La cinquième année, on relance un nouveau cycle de quatre ans ; les personnes enquêtées la première année, sont ré - enquêtées. Les entrants (naissances, immigrés) sont automatiquement introduits dans l'échantillon, les sortants (décès, émigrés) sont en principe exclus. Depuis 1998, ESPS est effectuée sur la moitié de l'EPAS tous les deux ans. L'échantillon de l'EPAS est donc enquêté tous les 4 ans, un quart tous les ans entre 1988 et 1997, une moitié tous les 2 ans depuis 1998 : entre1988 et 2002, cet échantillon a donc été enquêté à 4 reprises (1988-91, 1992-95, 1996-98, 2000-2002) et un cinquième cycle a démarré en 2004.

3.1. Les unités de sondage, grappe assuré et ménage.

18 *elles bénéficient d'une délégation de gestion de la Sécurité sociale et remboursent leurs adhérents aussi bien pour les prestations de Sécurité sociale que pour leurs prestations propres*

19 *Certaines données de Sécurité sociales sont transmises au fisc pour établir (ou contrôler) les revenus des personnels soignants libéraux*



Les unités de sondage sont différentes dans les deux sources d'observation. Dans EPAS, l'unité de sondage est la « grappe assuré », groupant l'assuré retenu à partir de la condition **C** et ses ayants droit, qu'ils appartiennent au même ménage que lui ou non (hachuré ascendant dans le schéma 1). Dans ESPS, cet assuré est dit **assuré principal** ; à partir de son adresse, on relève les informations sur tous les membres du ménage, qu'ils soient ses ayants droit ou non (hachuré descendant) ; l'intersection entre les deux sources est l'assuré principal et ses ayants droit cohabitants (croisillons).

Dans l'EPAS, certains ayants droit de l'assuré principal peuvent résider ailleurs que dans le ménage, en particulier dans le cas de ménages recomposés (partie en beige) , dans ESPS, certains membres du ménage peuvent être assurés à titre personnel ou ayants droit d'autres assurés que l'assuré principal, ces assurés pouvant appartenir eux même ou non au ménage (partie en bleu).

Dans l'exploitation des données de l'EPAS et de l'ESPS, l'unité d'analyse est toujours la personne, appartenant à la « grappe assuré » ou au ménage de l'assuré principal [5].

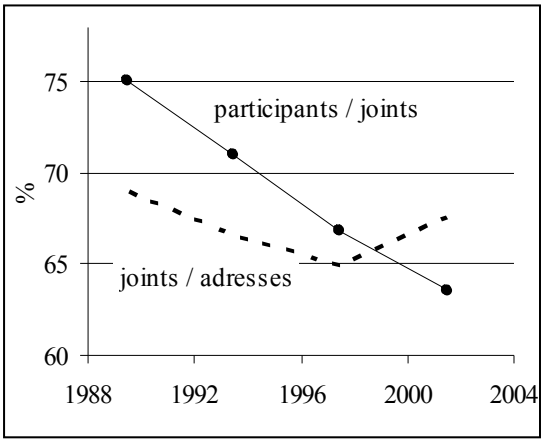
Notons que l'enquête est actuellement faite uniquement auprès des assurés résidant dans des « ménages ordinaires » et des autres personnes résidant dans ces ménages, qu'elles soit ayant droit des premiers ou non.

Nous désirions avoir également des informations sur les personnes résidant en institution ; elles présentent en effet un grand intérêt du fait de leur état de santé dégradé nécessitant des soins importants. L'enquête a donc également été faite sur cette population pendant les 4 premières années d'ESPS, de 1988 à 1991. Lorsque l'institution a accepté de répondre, le taux de réponse « personne » n'était que de 57 %, très différent selon la nature de l'institution, **38 %** pour les institutions très médicalisées (long séjour ou section de cure médicale), **64 %** pour celles peu médicalisées (maison de retraite ou logement foyer). En conséquence, à partir de 1992, l'ESPS s'est limitée aux ménages ordinaires. Nous enregistrons malheureusement dans ce cas un échec conduisant à une lacune dans la représentativité de l'échantillon potentiellement apparié[20].

Dans un premier temps, il avait été envisagé de recueillir auprès des ménages les seules données sociodémographiques (profession, niveau d'instruction, revenu, protection maladie complémentaire,…) et de morbidité (maladies, handicaps,…) ; en fin de compte, la décision a été prise de recueillir également des données de consommation médicale. Deux éléments en effet paraissaient manquer dans les données de Sécurité sociale, les consommations non remboursées et certains dépassements, et surtout on n'avait pas la certitude que l'importance de ces lacunes seraient uniformes selon les catégories sociales et stables au cours du temps : à ne mesurer la consommation médicale qu'à partir des remboursements, on se prive de la possibilité d'en connaître les évolutions en cas de modification de la part remboursée par la Sécurité sociale ou de non parallélisme des prix et des tarifs (augmentation ou diminution des dépassements).

20 *Ces personnes appartiennent l'EPAS, sans qu'on sache qu'elles résident en institution*

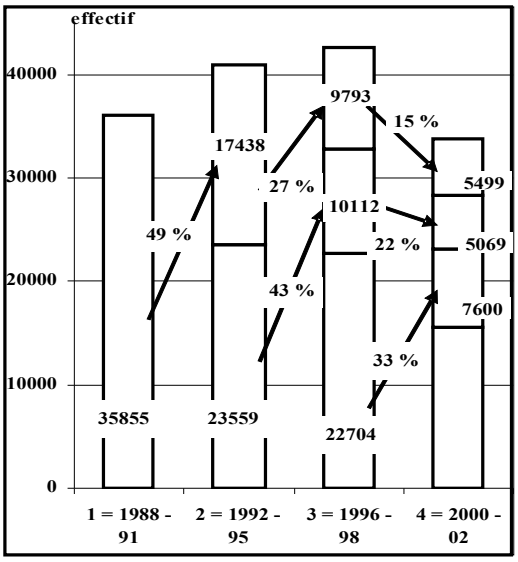
3.2. Les taux de participation



La base de sondage est constituée de l'EPAS, lui-même sondage dans les fichiers des 3 grands régimes de Sécurité sociale. Pour différents motifs (adresse incomplète ou erronée, logement vide, déménagement, personne âgée résidant en institution, décès non encore inscrit dans les fichiers de Sécurité sociale[21],…), le contact avec l'assuré ne peut avoir lieu. Le taux de ménages non joints se situe entre 30 % et 35 %.

Une fois le ménage joint, les taux d'acceptation donnent une indication de la bonne acceptation de l'enquête : le taux de participants (parmi les assurés joints) a régulièrement diminué depuis le premier passage, passant de 75,0 % pour le premier passage (1988-91) à 70,9 % (1992-95), 66,8 % (1996-98), 63,5 % (2000-02). Cet effet d'usure de l'échantillon est bien connu des gestionnaires de panels et devrait conduire à renouveler l'échantillon. Un tel renouvellement sera possible sans grande difficulté, puisque l'EPAS a été doublé, il suffira de prendre l'autre moitié. On perdra alors pour un temps l'intérêt de panel de cette enquête, puisque l'histoire de chaque personne repartira à 0.

3.3. Le chaînage



Le chaînage permet de relier deux ou plusieurs observations de chaque enquêté et représente de ce point de vue un avantage du panel sur une succession d'enquêtes transversales à quatre ans d'intervalle.

Parmi les enquêtés du premier passage, 49 % ont été retrouvés lors du deuxième passage, 27 % lors du troisième passage, 15 % lors du quatrième[22] passage. Rappelons que deux passages consécutifs chez un même enquêté sont espacés en général[23] de quatre ans.

Parmi les nouveaux enquêtés du deuxième passage, 43 % ont été retrouvés lors du troisième passage, 22 % lors du quatrième passage.

Parmi les nouveaux enquêtés du troisième passage, 33 % ont été retrouvés lors du quatrième passage.

Mais on rencontre d'autres configurations, certains enquêtés participant lors d'un passage peuvent refuser au passage suivant et accepter à un passage ultérieur.

Ainsi, parmi les enquêtés (enquête ménage) au premier passage, 11 % refusent lors du deuxième passage et acceptent lors du troisième, parmi eux, seuls 5 % participent au quatrième passage ; de même, parmi les nouveaux enquêtés du deuxième passage, 11 % refusent lors du troisième passage et acceptent lors du quatrième ; 3 % enfin des enquêtés du premier passage refusent de participer aux deuxième et troisième passage et acceptent au quatrième.

Soulignons que ces taux ne sont pas très élevés du fait de l'organisation des fichiers de Sécurité sociale et de la méthode d'enquête : le chaînage est fait sur tous les membres du ménage et pas seulement sur les seuls assurés principaux ou les seuls membres de la « grappe assuré » ; les enfants disparaissent du dossier de leurs parents lorsqu'ils deviennent autonomes ; de même sont enquêtées les personnes assurées à titre personnel résidant avec

21 *en cas de décès, la Sécurité sociale met ses fichiers à jour avec plusieurs mois de retard ou de déménagement (car les remboursements peuvent être demandés après de longs délais) ; en cas de déménagement, les assurés ne préviennent la Sécurité sociale que lorsqu'ils demandent une prestation, soit parfois beaucoup plus tard*

22

23 *lors de la modification du rythme de l'enquête, d'annuelle à bisannuelle, certains espacements ont pu être de 3 ans.*

les assurés principaux, si lors d'un passage ultérieur, cette co-résidence disparaît, elles sortent du champ de l'enquête. Ce motif de non chaînage devrait disparaître avec la future organisation des fichiers de Sécurité sociale qui supprimera les notions d'assuré et d'ayant droit, chaque personne sera assurée dès sa naissance sous son propre NNI Ces deux motifs de non chaînage viennent s'ajouter aux pertes « naturelles », décès et entrées en institution, l'enquête n'étant faite qu'auprès des ménages ordinaires.

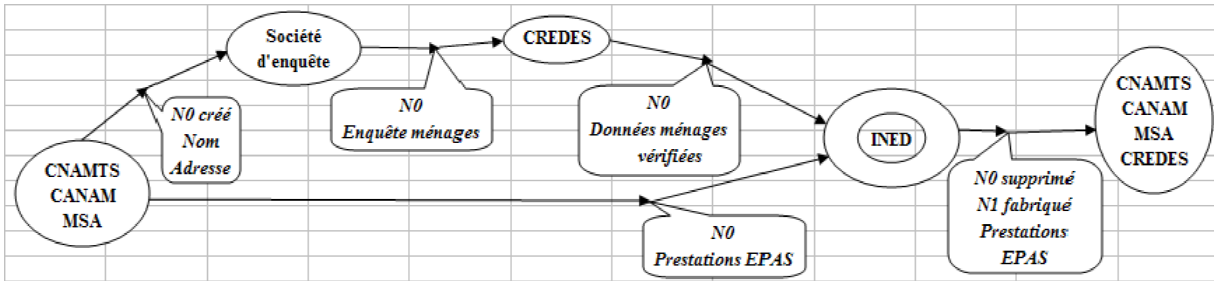
4. L'APPARIEMENT, SES LIMITES

Chaque ménage d'ESPS comprend au moins un assuré, celui de l'EPAS, dit **assuré principal** ; il peut comprendre d'autres assurés, dits assurés non principaux. Seuls l'assuré principal et ses ayants droit peuvent être appariés. Il faut regrouper les doublons, séparer les décès et les naissances en cours d'année (on n'a leur consommation médicale que sur une partie de l'année), et les adhérents des « mutuelles décompteuses[24] ». On retire aussi les personnes se déclarant assurées (non principales) et les ayants droit des assurés non principaux n'ayant eu aucune prestation. En 2000, sur 20 045 personnes enquêtées, 11 257 étaient en principe appariables (56,2 %).

On retient 6 critères d'appariement : NNI de l'assuré, 4 premières lettres du prénom, sexe, jour, mois et année de naissance. Si on tolère une erreur sur l'un de ces 6 critères, on apparie 88 % des personnes appariables ; on en ajoute 3 % si on remplace les 4 premières lettres du prénom par la première lettre seulement, et encore 2 % si on abandonne le prénom (appariement 2000). Dans ces conditions, on apparie 93 % des personnes appariables (12 025). En 2000, on aboutit ainsi à 9 116 personnes satisfaisant rigoureusement aux critères de sondage de l'EPAS (81 %), résidant dans un ménage ordinaire, en France, protégées par le régime général (à l'exception des ressortissants des mutuelles décompteuses), des agricoles ou des indépendants, ayant accepté de participer à l'ESPS, dont les informations d'identification sont cohérentes dans les deux sources.

L'appariement porte sur les personnes, non sur les consommations médicales.

La procédure d'anonymisation



L'anonymisation se fait en double aveugle par la création d'un double circuit et par l'intervention d'un tiers non impliqué, l'INED :

- premier circuit ou circuit long : les trois régimes de Sécurité sociale, CNAMTS, CANAM, MSA, après avoir supprimé le NNI, envoient à une société d'enquête le nom et l'adresse des assurés de l'EPAS, ainsi qu'un numéro N0, créé pour l'appariement ; à partir des coordonnées des enquêtés, la société d'enquête fait l'enquête auprès des ménages, et envoie les résultats au CREDES après avoir supprimé les identifiants, nom et adresse ; le CREDES vérifie les données obtenues avant de les envoyer à l'INED,
- deuxième circuit ou circuit court : les trois régimes de Sécurité sociale, CNAMTS, CANAM, MSA, après avoir supprimé le NNI, envoient à l'INED les données de prestations, identifiées par N0,
- appariement : l'INED supprime N0 et le remplace par un nouveau numéro N1. On peut ainsi apparier les données de Sécurité sociale et les données d'enquête sous le numéro N1 ; ces données sont envoyées ainsi rendues anonymes à la CNAMTS, la CANAM, la MSA et au CREDES pour exploitation.

Aucun organisme n'a, à aucun moment, en même temps les prestations, les données d'enquête et les moyens d'identifier les personnes. La CNIL[25] a jugé cette procédure fiable et a donné l'autorisation de la mettre en œuvre.

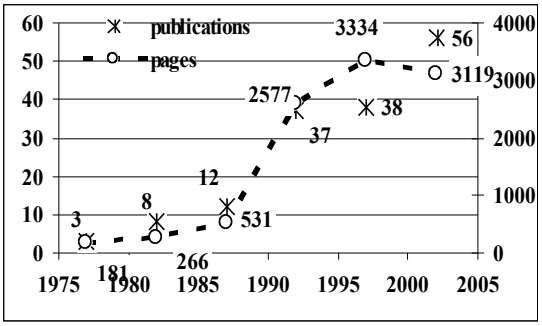
L'appariement entre les données de l'EPAS et celles d'ESPS a été faite pour 1988, 1992, 1995 1997 1998, 2000, il est fait en routine depuis 2000.

zzz

24 *Cf. partie 1*

25 *La Commission nationale sur l'informatique et les libertés délivre les autorisations de création des fichiers informatiques concernant les informations sensibles sur les personnes*

5. CONCLUSION, LES PUBLICATIONS, LES ÉVOLUTIONS



La masse d'information recueillie est considérable et a donné lieu à un nombre de nombreuses publications considérable et qui va en croissant. Les travaux ont d'abord été seulement ceux des statisticiens de la Sécurité sociale, puis du CREDES (aujourd'hui IRDES) ; depuis quelques années, les statisticiens du Ministère de la santé ont commencé à les utiliser, et plus récemment, les chercheurs universitaires et des autres centres de recherche.

De nombreux thèmes n'ont pas encore été traités du fait de la richesse de l'information recueillie et de sa complexité ; en particulier, peu de travaux longitudinaux au niveau individuel.

De nombreux et importants changements sont intervenus depuis les premiers sondages, qui ont d'importantes répercussions sur la manière d'aborder les problèmes : en premier lieu l'introduction et diffusion généralisée de l'informatique à tous les stades, aussi bien dans le recueil des données auprès des ménages (enquête ménage assistée par ordinateur) , et dans leur saisie que dans l'augmentation des capacités de stockage et dans la plus grande facilité du traitement des données. En même temps, nous avons assisté à une extension considérable du rôle de la statistique (aide à la décision, base de négociation, …) et de son acceptation auprès de nos interlocuteurs ; cette extension de la statistique s'est accompagnée d'une harmonisation des nomenclatures. Egalement, l'introduction des législations de protection des données personnelles, en même temps qu'elle compliquait notre tache, donnait un cadre plus formalisé et plus général à nos préoccupations de déontologie professionnelle. Plus généralement, l'extension de la population couverte par la Sécurité sociale donne un intérêt plus grand aux données qu'elle crée au cours de son activité.

Lorsque nous avons commencé à travailler sur ce projet, nous n'avions conscience pas des difficultés qui nous attendaient, délais d'autorisation et de mise en œuvre des recueils, et appariements entre données de natures différentes. En effet, les fichiers de Sécurité sociale sont adaptés à la gestion des remboursements, les transformer en fichiers utilisables pour la recherche et les apparier avec une enquête ménage est difficile et souvent un peu décevant (le nombre d'erreurs est important et on n'en a pas conscience lorsqu'on ne dispose que d'une seule source.

Il paraît intéressant de conclure sur les inconvénients et les avantages des sondages par rapport aux données exhaustives, qu'il s'agisse de données de population (recensement versus sondage) ou de données de nature administratives telles les données de Sécurité sociale (cela doit être également vrai pour les données industrielles ou commerciales).

Du côté des inconvénients, la précision la précision des estimateurs des paramètres est moindre dans le sondage que la « valeur réelle » obtenue à partir du recensement ; la taille de l'échantillon peut en particulier être insuffisante pour des recherches pointues sur des cas rares. La solution peut être trouvée dans l'augmentation des taux de sondage, en évitant toutefois d'approcher des échantillons proches de l'ensemble de la population.

Du côté des avantages, on a d'abord un moindre coût, ce qui permet d'affecter une partie des dépenses au contrôle et à l'amélioration de la qualité des données ; nous pensons ce problème de la qualité des données est tout à fait essentiel, car comment espérer des conclusions convenables à partir d'informations parfois gravement entachées d'erreurs. Le moindre coût permet également de prévoir un sondage permanent dont l'intérêt est considérable : rapidité des premiers résultats, disponibilité des données, souplesse par l'introduction de modules mobiles. Enfin, du fait qu'une toute petite partie de la population est enquêtée, le sondage évite le risque ou de la tentation de contrôler la population ou d'utilisation déviante de données personnelle.

Petit lexique

Sécurité sociale : Assurance maladie obligatoire

Régimes de Sécurité sociale : la Sécurité sociale est géré de manière autonome par des caisses selon le statut professionnel : régime général (salariés du commerce et de l'industrie) et rattachés (fonctionnaires, étudiants, artistes et auteurs, médecins et auxiliaires conventionnés, CMU,…), régimes des indépendants (artisans, commerçants, libéraux), régime des agricoles (exploitants et salariés), régimes dits spéciaux (SNCF, EDF-GDF, mineurs, marins pêcheurs, militaires,…).

CNAMTS : Caisse nationale d'assurance maladie des travailleurs salariés (régime général)

CRAM : Caisse régionale d'assurance maladie (régime général)

CPAM : Caisse primaire d'assurance maladie (régime général, en général, une par département)

CANAM : Caisse nationale d'assurance maladie des professions indépendantes, (artisans, commerçants, libéraux)

MSA : Mutualité sociale agricole (exploitants et salariés)

SNCF, EDF-GDF : Société nationale des chemins de fer, Electricité de France – Gaz de France

Consommant (au sens de la Sécurité sociale) : personne ayant entraîné une dépense de l'assurance maladie pour des soins médicaux au cours de l'année

Prestataire : personne ayant reçu une prestation de l'assurance maladie au cours de l'année

Assuré ou Immatriculé (à un régime de Sécurité sociale) : personne ayant fourni les documents ouvrant ses droits à être recevoir les prestations de ce régime

Ayant droit : personne dépendant d'un assuré (enfant, épouse ou ascendant) pour l'ouverture des droits à l'Assurance maladie

Personne protégée : assuré ou ayant droit

Grappe assuré : ensemble d'un assuré et de ses ayants droit

Liquider un dossier, liquidateur : gérer un dossier de remboursement et ordonner le versement des prestations, personne faisant ces opérations

Mutuelle décompteuse : mutuelles habilitées à gérer les prestations de l'assurance maladie obligatoire d'une catégorie professionnelle (enseignant, employés des postes,…) et qui reçoivent pour cela une ristourne de gestion

Couverture complémentaire : la Sécurité sociale rembourse une part de plus en plus faible des consommations médicales ; pour le reste, une proportion de plus en plus importante de la population recourre à une assurance complémentaire, mutuelle (code de la mutualité), assurance privée (code des assurances) ou caisse de prévoyance

NNI : numéro national d'identité, donné par l'INSEE à chaque personne à sa naissance

CNIL : comité national de l'informatique et des libertés

Bibliographie

[1] Anonyme, (Jalibert ?) (1962), « Une enquête sur 400 000 cas de maladie, résultats d'ensemble », in FNOSS, *Notes et documents N° 8*.

[2] Anonyme, (Sanson-Carette, A. ?, Nathan, R. ?) (1969) *Résultats de l'enquête statistique par sondage de la structure des frais médicaux, Notes N° 1, 2, 3*, Paris : Ministère des affaires sociales, Direction de l'assurance maladie et des caisses de Sécurité sociale, Sous direction des affaires financières,.

[3] Bocognano, A., (1990), *Méthode et déroulement de l'enquête sur la santé et la protection sociale.* Paris : CREDES, 131 pages.

[4] Dumesnil, S., Lebreton, S., (1995), *Comparaison méthodologique des enquêtes : Santé et soins médicaux 1991-1992 - Santé et protection sociale 1992.* Paris : CREDES, 209 pages.

[5] Grandfils, N., (1994), *Prestations, santé et protection sociale : une approche socio-économique : première exploitation de deux enquêtes jumelées : EPAS - ESPS 1988.* Paris : CREDES, 246 pages.

[6] Guillot, G., Mizrahi, An, Mizrahi, Ar, (1968), « Etude critique de méthodes d'enquête (une enquête pilote sur la consommation médicale) », in *Consommation, n° 1*. pp. 3-38.

[7] Mizrahi, An, Mizrahi, Ar, (1970), *Enquête nationale sur les soins médicaux 1970. Présentation de l'enquête auprès des ménages.* Paris : CREDOC.

[8] Mizrahi, An, Mizrahi, Ar, (1975), *L'enquête sur la santé des Canadiens. Aspects techniques de l'entrevue et appariement avec les fichiers d'Assurance Maladie et hospitalisation.* Paris : CREDOC, 55 pages.

[9] Mizrahi, An, Mizrahi, Ar, (1977), *Projet de sondage dans les fichiers de Sécurité Sociale.* Paris, CREDOC, 39 pages.

[10] Anonyme, (Gaillard, A., Nathan, R. ?) (1978), *PANEL, les premiers résultats*, Agence comptable, Département statistique, Paris : CNAMTS.

[11] Mizrahi, An, Mizrahi, Ar, (1978), *Méthode de sondage. Enquête permanente dans les dossiers de Sécurité Sociale.* Paris : CREDOC, 50 pages.

[12] Péquignot, H., Etienne, J.P., Parmentier-Lemoigne, N., (1960), « Etude sur les variations de la consommation médicale en fonction de l'âge et du sexe », *La semaine des hôpitaux, 36ème année, N° 9-10*, Paris, p. 228-233. Thèse de médecine de Parmentier-Lemoigne, N., Paris 1960,

[13] Volatier, J.-L., (1988), *Enquête sur la protection sociale. Résultats méthodologiques de l'enquête expérimentale.* Paris : CREDES, 104 pages.

*** *** *** *** ***