LES BESOINS DES ECONOMISTES EN MATIERE D'APPARIEMENTS SECURISES

Nathalie PICARD(*), Benoît RIANDEY(**), Anne SOLAZ(**)

(*)Université de Cergy-Pontpoise, INED et Ecole Polytechnique
(**) INED

Introduction

Les économètres peinent à la recherche de données longitudinales. Les panels sont rares en France malgré les avancées récentes appuyées par le rapport au CNIS de Stéfan Lollivier et Mylène Chaleix [4], notamment le changement d'échelle en cours de l'Echantillon démographique permanent (EDP). Aussi sont-ils souvent conduits à se limiter aux panels étrangers, notamment anglais BHPS (British Household Panel Survey), allemand GSOEP (German Socio-economic Panel) ou américains. Le panel PSID[1] (Panel Study of Income Dynamics) du Michigan est l'exemple type de sources qui apportent sur les ménages et pour une très longue durée (depuis 1968) « tout ce que les chercheurs désirent savoir ». Certes, on peut débattre de la représentativité à long terme de ces panels, mais la longueur autorise des études de long terme tant sur les carrières professionnelles et salariales que sur les parcours familiaux. Quant aux panels renouvelés par partie comme l'enquête Emploi en France, ils ne permettent de disposer sur un ménage que de données à moyen terme, six trimestres pour l'enquête emploi en continu. De plus, ces panels étrangers sont d'assez grosse taille pour permettre d'étudier des événements rares comme les conséquences du divorce ou du veuvage, la mobilité résidentielle de longue distance. Reproduire ce type d'études sur les parties françaises des panels européens (ECHP puis SILC) serait très vite limité par la petite taille des échantillons. Les chercheurs voudraient donc faire appel aux sources administratives**.** Nous allons proposer des solutions pour apparier données d'enquêtes ou de panels et sources administratives dans le respect de la confidentialité. De vastes perspectives s'ouvrent à la statistique publique si elle se mobilise sur ces méthodes avec l'audace suggérée par la directrice du service juridique de la CNIL [8].

1. Les panels de la statistique publique aujourd'hui

A l'heure actuelle, la statistique publique française dispose de trois[2] grands panels de méthodologies bien distinctes : l'enquête emploi, l'échantillon démographique permanent et le(s) panel(s) de l'assurance maladie[3]. Chacune de ces sources associe déjà des données statistiques et des données administratives (fiscales et maintenant de prestations sociales pour l'Enquête emploi, d'état civil pour l'EDP, de consommation médicale pour l'ESPS-EPAS). Pour l'enquête emploi et l'EDP, on enrichit une source statistique avec une source administrative. Pour l'enquête ESPS, on enrichit de questionnaires une source administrative. A contrario, les panels EDP et EPAS- ou EPIBAM sont des sources permanentes, continuellement mises à jour, contrairement à l'Enquête emploi. L'idée que nous allons développer dans cet article est la suivante : Ne pourrait-on envisager de pérenniser durablement l'échantillon de l'enquête emploi par le recueil de données administratives ?

---

[1] Pour plus d'informations sur le PSID, voir le site http://psidonline.isr.umich.edu/

[2] Nous ne citons pas les échantillons français des panels européens (actuellement SILC) en raison de leur petite dimension nationale.

[3] Le vieux panel EPAS (échantillon permanent des assurés sociaux) de la CNAMTS (régime général), de la Mutualité agricole et du régime des indépendants vient de donner naissance à un rejeton l'EPIB(AM) (Lenormand, [17]), l'échantillon permanent de bénéficiaires (de l'assurance maladie), un échantillon au 1/100ème des bénéficiaires de l'assurance maladie. Dans un terme non immédiat, l'ensemble des régimes d'assurance maladie sont appelés à participer à l'EPIBAM. L'enquête ESPS (enquête santé et protection sociale) de l'IRDES (ex-CREDES) est un panel (principalement) téléphonique fondé sur ces échantillons appariant les questionnaires et les données de consommation médicale de l'EPAS, bientôt de l'EPIBAM. Dans l'attente de son extension aux autres régimes, la CNAMTS nomme l'EPIBAM « échantillon général de bénéficiaires » (EGB).

1.1. Marier enquête et fichiers de l'emploi ?

Sur ce thème de l'emploi et plus précisément du chômage, les statisticiens publics opposent avec regret les qualités et limites de ces deux types de sources :

- seule l'Enquête emploi permet de mesurer un taux de chômage au sens du BIT (Bureau International du travail) ainsi que l'emploi dit 'inadéquat' et de mener des analyses socio-démographiques fines grâce à la richesse du questionnaire, mais l'analyse de l'emploi et du chômage au niveau local n'est pas accessible avec cette source. De plus, le suivi individuel dans le temps est effectué sur une durée trop courte pour calculer les droits à prestations ou analyser la dynamique de l'emploi au niveau individuel.

- les fichiers administratifs de façon continue, le recensement rénové à un rythme soutenu mais moindre, fournissent quant à eux des estimations au niveau local, par exemple par bassin d'emploi, mais ne peuvent en revanche prétendre à mesurer le taux de chômage au sens du BIT, ni à une analyse socio-démographique fine. Ils ne disposent pas en effet des variables et questions nécessaires.

Les statisticiens publics s'accommodent mal de ces limites (CNIS, *Chroniques* n° 8, [5]). Un anthropologue aurait immédiatement suggéré : « Hé bien, mariez les ! ». C'est ce qu'autoriserait un appariement sécurisé de l'enquête emploi et de fichiers administratifs de ce domaine. Examinons le potentiel de cette idée avant d'en critiquer la mise en œuvre.

La méthode Malinvaud de mesure infra-annuelle du taux de chômage s'appuie uniquement sur la confrontation des données statistiques et non celle des données individuelles. Lorsque, fin 2006, la série mensualisée s'est trouvée en décalage avec les résultats de la nouvelle enquête emploi, on aurait aimé pouvoir retourner aux données individuelles pour identifier les situations personnelles sources de cette divergence. C'est ce que permettrait l'appariement des données administratives relatives aux personnes enquêtées par l'enquête emploi avec les données des questionnaires de l'enquête, qu'il s'agisse des données DADS ou de l'ANPE. L'analyse des données administratives s'enrichirait de toutes les variables socio-démo-professionnelles de l'enquête emploi, ce qui éclairerait la compréhension des décalages entre les sources et les mesures du chômage et de l'activité selon les deux sources.

Au delà de l'intérêt méthodologique certain, cela permettrait aussi de répondre aux demandes locales grâce à la mise en œuvre des modèles d'estimation sur les petits domaines que Pascal Ardilly a introduits dans la statistique publique française [1], mais déjà bien développées dans d'autres instituts de statistique, notamment canadien. En témoigne le colloque de l'Association internationale de statisticiens d'enquête (AISE) à Riga en 1999 sur ce thème.

Donc finalement, les appariements sécurisés entre enquêtes et fichiers administratifs permettraient d'apporter aux fichiers administratifs les concepts mis en œuvre et la richesse informative des questionnaires d'enquêtes. Il est à noter que cette perspective n'avait pas été envisagée par le groupe de travail du CNIS présidé par Jean-Baptiste de Foucauld, sans doute parce que, en France, nous en sommes encore à la phase d'expérimentation méthodologique[4].

La comparaison des fichiers administratifs et d'enquêtes peut être utile en soi, mais elle ouvre aussi des perspectives. Si des données contemporaines de l'Unedic et des DADS étaient collectées auprès des enquêtés de l'enquête emploi, pourquoi ne pas continuer à les recueillir et à les apparier au fichier en aval, c'est à dire après le sixième trimestre quand le ménage est sorti de la collecte par questionnaire de l'enquête emploi ? Pourquoi ne pas aussi recueillir et apparier ces données de manière rétrospective sur une période de quelques années avant l'enquête emploi ? L'enquête emploi deviendrait alors une sorte de panel administratif permanent qui, qui plus est, s'accroîtrait de 9.000 ménages par trimestre (18.000 après doublement de l'échantillon).

[4] Remarquons qu'il n'y a pas de session, ni de communication à ces JMS sur les estimations sur petits domaines, mais des travaux sont envisagés à Lyon.

Une telle perspective peut laisser dubitatifs des statisticiens d'un institut peu familier des panels de longue durée. Une expérimentation pourrait d'abord être menée sur une vague de sortants, ce qui éclairerait immédiatement les divergences entre sources, puis permettrait d'envisager une extension de l'opération et éventuellement le prolongement post-enquête[5]. Reste à débattre de la question juridique et technique de l'appariement. Là encore les épidémiologistes et les économistes de la santé disposent d'une expérience qui nous serait profitable.

L'usage général du Nir dans la sphère sociale ouvre en théorie une très large palette de données appariables, mais les statisticiens n'abordent une telle perspective qu'avec la plus grande prudence. Les anciens sont encore marqués par les affres de l'affaire SAFARI[6]. Les plus jeunes hésitent devant sa complexité administrative car l'exigence protectrice d'un décret en Conseil d'Etat[7] a longtemps limité ces appariements à quelques opérations phares, comme le vieux panel DAS, les échantillons permanents de retraités ou de cotisants ou le panel des bénéficiaires de minima sociaux. Les techniques anonymes d'appariements sécurisés[8] (par cryptage irréversible, le hachage) évitent cette lourde démarche institutionnelle. Elles permettraient d'envisager le suivi anonyme de carrières individuelles en rapprochant les données des DADS et celles de l'UNEDIC ou de l'ANPE. Ainsi les économistes étofferaient leurs analyses de la dynamique de l'emploi et de la recherche d'emploi. Plusieurs pistes sont envisageables ; au cours de cette cession, Catherine Quantin [20],[22] en propose une qui renouvellerait profondément l'approche de la statistique publique.

1.2. Les médecins marient leurs cohortes aux fichiers d'emploi

Les enquêtes santé 2002 et 2008 viennent de renouer avec l'appariement des données de sécurité sociale (via le Nir) expérimenté en 1970 pour l'enquête santé décennale INSEE-CREDOC [18], mais longtemps suspendu par la lente digestion de la loi Informatique et Libertés de 1978.

Aujourd'hui l'INSERM et l'assurance maladie mettent en œuvre la Cohorte Constances [9] de 200.000 patients et la plate-forme Plastico destinée à en gérer les appariements entre diagnostics dans les centres de santé, consommation médicale (SNIIRAM-PMSI), carrières (CNAV), état civil (INSEE) et causes de décès (INSERM Cépidc) ; autant dire un projet autrement sensible et ambitieux que les quelques lignes précédentes relatives aux appariements dans le domaine de l'emploi. Depuis ce passionnant article du *Courrier des statistiques*, le projet s'étoffe d'une expérimentation très féconde : la première application de Plastico sera en 2009 l'appariement de toutes ces données complémentaires avec non pas le futur Constances, mais le bien présent EPIBAM, et donc à terme avec l'enquête ESPS de l'IRDES. Il en sera fait état au prochain Séminaire Appariements sécurisés de la SFdS[9] qui se tiendra à Paris à son siège le 16 novembre prochain.

1.3. Engager une réflexion collective

Les épidémiologistes sont confrontés au besoin de rapprocher les diagnostics et les dépenses de santé dans le respect du secret médical. Ils réfléchissent donc avec la CNIL à leur accès au Nir, sous un de ses avatars, avec toutes les garanties apportées par leur finalité à la fois médicale, statistique et scientifique. C'est d'ailleurs avec eux qu'en 1999, à l'initiative de la Commission de déontologie de la SFdS, que le CNIS, l'INSEE et d'autres chercheurs ont réfléchi à la transposition de la directive européenne du 24 octobre 1995 [6] et obtenu le 6 août 2004 par cette transcription, la reconnaissance des garanties apportées par les finalités de statistique et de recherche scientifique ou historique

[5] Nous examinerons plus loin les difficultés de la panélisation de l'unité ménage.
[6] L'opération d'informatisation du RNIPP avait maladroitement été nommée Safari, d'où le succès de l'article du journal *Le Monde* « Safari ou la chasse aux Français ». Cet article fut à l'origine positive de la loi Informatique et Libertés et négative d'un blocage durable de l'usage du Nir et des appariements pour la statistique publique.
[7] Article 18 de la loi Informatique et libertés initiale ; article 27 de la loi modifiée le 6 août 2004.
[8] Avant même le séminaire Appariements sécurisés de l'INED et de la SFdS de février 2001, la première communication de Catherine Quantin proposant le transfert de ces méthodes de l'épidémiologie à la statistique publique date des Journées de méthodologie statistique de décembre 2000 [20]; les premières applications à la statistique publique (loi de 1951) par Alain Goy et la DEPP datent de 2004 [10] ; leur enseignement à l'ENSAI interviendra pour la première année au printemps 2009.
[9] Depuis 2006, le cours ou le séminaire Appariements sécurisés de la SFdS se tiennent tous les 16 novembre œuvrés (et donc par nécessité le 13 novembre en 2008 et en 2013 et 2014).

(article 6 de la loi modifiée) et la dispense pour l'INSEE et les SSM de l'accord exprès pour le recueil des données sensibles sur avis du CNIS (article 8, alinéa II.7)[10]. Ce groupe de travail s'est penché sur l'accès aux bases de sondages, mais il n'y fut jamais question de l'usage du Nir. C'est la CNIL qui prit l'initiative d'infléchir l'article 18 en l'article 27 pour tenir compte des usages sécurisés du Nir, affirmant régulièrement que le Nir haché n'est pas le Nir et ne relève donc pas de l'obligation d'un décret en Conseil d'Etat.

L'évocation de ce groupe de travail, déconnecté de tout dossier d'actualité présenté à la CNIL, montre la fécondité d'une telle réflexion. Ne serait-il pas opportun de renouveler cette expérience dans le cadre des appariements sécurisés et du traitement des identifiants ? Réunir à nouveau ce groupe de travail permettrait à la statistique publique de concevoir un nouveau système d'accès sécurisé et communiquant des fichiers administratifs. L'enjeu mérite cette peine. L'article du *Courrier des statistiques* de mai 2007 [8] apporte une vision renouvelée de la qualité et des mises en connexion des identifiants sectoriels[11]. Quand, au début des années 80, la CNIL a imposé la sectorisation des identifiants, aucune solution n'a été apportée à leur validation ni à l'éventualité d'échanges entre secteurs. Les techniques actuelles permettent d'apporter une cohérence globale à cette politique.

Les statisticiens ont conscience de l'intérêt public très fort des statistiques relatives à l'emploi et au chômage et des statistiques locales. L'association nouvelle apportée par cette communication, entre appariements sécurisés d'enquêtes et de fichiers administratifs et estimations locales sur des petits domaines leur apporte une légitimité très forte pour aborder avec une CNIL, fort bienveillante, les questions techniques et organisationnelles qui permettraient cette avancée d'une très grande généralité. On peut d'ailleurs regretter que l'enquête santé 2002, qui associait appariement avec le SNIIR-AM et extensions régionales n'ait pas été l'occasion de tester cette méthodologie d'extension aux données locales des résultats nationaux de l'enquête. Les extensions régionales sont un luxe, difficilement généralisables. Leur utilisation méthodologique dans le sens proposé serait d'un grand bénéfice collectif.

1.4. Une nouvelle architecture des appariements

Catherine Quantin lève une limitation grave des appariements sécurisés. Dans leur finalité protectrice, le hachage des identifiants ne permet que les appariements préalablement définis à la mise en œuvre du projet. C'est ce qu'impose le principe de finalité de la loi de 1978, mais c'est exclure toute extension de finalité, pourtant bien envisagée par la révision de la loi en 2004. Comment ne pas exclure techniquement un traitement ultérieur qui aurait l'aval de la CNIL ?

Les épidémiologistes ont toujours été conscients de l'obligation éthique de pouvoir lever l'anonymat si un patient requiert des soins au vu d'un diagnostic mis en œuvre par la recherche. Il n'y a donc pas de procédure épidémiologique anonyme sans tiers de confiance pouvant partiellement lever l'anonymat.

Pour la seule finalité de recherche, les épidémiologistes peuvent souhaiter apparier des fichiers de recherche initialement autonomes (par exemple une recherche exhaustive sur le diabète dans les fichiers d'assurance maladie avec une recherche sur une maladie aggravée par le diabète). La solution proposée par Catherine Quantin et approuvée par la profession [22], [23]) est la suivante : l'identifiant santé (éventuellement le Nir) est haché (de façon irréversible) dans toutes les études épidémiologiques avec la même clé, puis est crypté (de façon réversible) avec une clé spécifique à chaque recherche conservée par une autorité de gestion des clés émanant de la CNIL. Pour un appariement autorisé, le déchiffrage des identifiants permet de revenir à l'identifiant haché pour constituer le fichier joint.

Cette architecture est envisageable pour la statistique publique à une grande échelle. Certes, il ne s'agit pas de refuser le maintien dans leur technique actuelle des opérations menées à l'aide du Nir en clair. Il n'y a pas lieu de compliquer ce qui est autorisé. Mais on pourrait décider que les fichiers

[10] La communauté statistique doit être particulièrement reconnaissante à Gérard Lang pour avoir porté ces demandes au cours des longues négociations interministérielles.

[11] Pour seulement dédoublonner l'identifiant fiscal, il n'aurait pas été nécessaire de réintroduire le Nir, mais seulement un Nir haché ; c'est ce dont l'Ine de l'Education national aurait besoin ; par contre, pour mettre en relation, ces différents secteurs, le Nir est utile, à moins de lui substituer le dispositif présenté aujourd'hui même par Catherine Quantin et évoqué ci-après.

administratifs comprenant le Nir soient désormais toujours transmis à la statistique publique muni d'un Nir haché (avec la clé de la statistique publique), puis crypté (avec une clé spécifique).

Si on met en place la validation des identifiants sectoriels (tel l'Ine de l'Education nationale) à l'aide du Nir haché (comme proposé en 2007 [8]), cette architecture sécurisée s'étend à l'ensemble des identifiants sectoriels. Bien sûr, la sécurité d'un identifiant haché n'apporte aucune garantie quant au caractère potentiellement indirectement nominatif du fichier par recoupement entre variables. Cette vigilance reste nécessaire.

Ce développement de portée générale invite les statisticiens et le CNIS a d'abord s'assurer de ce qu'ils veulent améliorer dans leur diffusion statistique avant de s'arrêter sur les contraintes techniques qui peuvent être élargies dans le nouveau cadre proposé.

Bien évidemment la technique ne suffit pour tout réussir. La statistique suppose la confiance, mais cette crainte de perdre la confiance ne doit pas être un facteur de blocage immédiat opposé aux tentatives d'innover.

Ces perspectives techniques ouvertes, il est temps de revenir aux besoins des économistes.

2. Articuler les thèmes au niveau du ménage

Il est assez facile d'imaginer l'enrichissement de panels d'individus pour les traitements précédents à l'aide de données administratives anonymisées. Mais il semble que ce soit plus difficile en France quand on élargit la recherche à l'entité du couple, contrairement aux pays disposant de registres de population[12]. Le couple est une entité qui varie au cours du temps, qui peut se former puis se dissoudre. Etudier par exemple les trajectoires professionnelles de couples n'a été possible en France qu'à partir des enquêtes disposant d'un calendrier retraçant de manière rétrospective l'histoire professionnelle et conjugale comme dans l'enquête Jeunes et carrières 1997 ou Familles et Employeurs 2005. Mais ces enquêtes ont trois inconvénients. Les effets de biais liés aux problèmes de mémoire sont bien connus. Le second inconvénient quand on travaille sur les couples est que l'on ne dispose au moment de l'enquête que des couples stables (n'ayant pas connu de séparation à ce jour), ce qui limite par exemple les études sur les conséquences de la séparation puisqu'on ne peut observer qu'un des partenaires. Enfin, l'impossibilité d'obtenir des données très précises comme les salaires par exemple en rétrospectif est une limite importante pour la richesse des études. La récente volonté d'ouvrir le champ des utilisateurs des fichiers fiscaux est une piste intéressante, même si la définition du ménage fiscal reste différente de celle du couple conjugal pour les couples non mariés et non pacsés qui restent très nombreux en France. Néanmoins, n'utiliser que les seuls fichiers administratifs exhaustifs reste insuffisant pour observer les biographies professionnelles de couples.

Si l'échantillon démographique permanent révèle la composition du ménage, il semble inadapté à s'enrichir de la biographie professionnelle du conjoint. La requête des économistes n'est donc pas simple à satisfaire.

L'enquête emploi en continu semble donc seule à répondre à ce besoin, mais seulement dans une perspective de court terme (six semestres), insuffisante pour appréhender une mesure longitudinale des transferts. C'est la raison pour laquelle nous suggérons de la prolonger par des années de carrières observées dans les fichiers administratifs comme nous l'avons proposé ci-devant. Cette proposition ne prend pas en compte la dynamique de la composition du ménage. Faute de registre de population et d'exhaustivité du recensement, la source fiscale semble la seule source administrative exhaustive qui puisse rendre compte des revenus sur plusieurs années à la réserve près avancée ci – dessus sur les couples non mariés et non pacsés.

Néanmoins si le sixième sortant est apparié avec les données contemporaines d'emploi ou fiscales, ne pourrait-on récupérer rétrospectivement les années antérieures ? Disposer par exemple d'un description des situations professionnelles dans les cinq années qui précèdent est une information cruciale pour les économistes afin de simuler les prestations publiques auquel l'individu pourrait

[12] Voir le site de l'INSEE pour une synthèse des types de recensement de populations http://www.insee.fr/fr/publics/default.asp?page=communication/recensement/particuliers/etranger.htm

prétendre au moment où il est interrogé lors de l'enquête Emploi. Cette information constituerait déjà une bonne base d'analyse des transferts en direction des familles.

Trop souvent les économistes ne trouvent pas leur compte auprès des enquêtes socio-démographiques. Ils en déplorent fréquemment l'absence des données économiques essentielles comme le revenu, ou la présence de données économiques de faible qualité. Pourtant, des données administratives de qualité seraient mobilisables, notamment celles des DADS. Les données fiscales, d'abord mobilisées par le feu CERC, viennent maintenant enrichir les enquêtes de l'INSEE. Leur usage pourrait être beaucoup plus large comme le mentionnait Jean-Michel Charpin, moyennant le recours aux plus strictes méthodes de la confidentialité.

Composition du ménage, emplois, ressources, voilà des données essentielles réclamées par les économistes, mais ils en réclament encore bien d'autres et les utilisateurs du PSID les ont obtenues au cours des vagues successives de collecte pendant des décennies. Ecoutons les :

Les économistes aimeraient pouvoir travailler sur des vrais panels longs avec des tailles d'échantillon suffisants permettant d'observer les mêmes individus suivis dans le temps :

* l'offre de travail (salaire, durée hebdomadaire) des différents membres d'un ménage, avec suivi des conjoints en cas de séparation et des "nouveaux" conjoints en cas de remise en couple après constitution du panel initial, avec continuité d'observation en cas de non emploi (chômage, inactivité) ;

* l'histoire démographique de l'individu tout au long de son cycle de vie, avec des renseignements sur ses conjoints successifs, ses enfants, leur éducation, etc (ces enfants étant par la suite suivis sur le marché du travail) ;

* l'historique des différentes résidences des individus,

* l'historique du patrimoine individuel/familial ;

* des panels de consommation avec biens exclusifs (consommés par un seul membre du ménage) et/ou assignables (dont on peut observer la quantité consommée par chacun des membres du ménage).

L'idéal, si l'on s'autorise à imaginer un monde idéal du point de vue des statisticiens et des économètres, serait évidemment d'avoir toutes ces informations pour les mêmes individus, afin de pouvoir les combiner pour analyser les interactions entre les choix correspondant à chacun des cas cités. Mais, avec moins d'exigences et plus de réalisme, pouvoir déjà relier deux de ces domaines constitue un pas en avant. Car c'est aujourd'hui l'interaction des domaines par exemple professionnels et familiaux, familiaux et résidentiels qui permet de traiter des sujets originaux et de disposer d'instrument économétriques adéquats. Les enquêtes spécialisées ont tout à gagner de s'enrichir d'autres sources de données.

Les thématiques de recherche liées pourraient ressembler à ce qui suit (extrait de différents projets de recherche) :

Les individus effectuent tout au long de leur cycle de vie des choix ayant des implications à court, moyen et long termes à la fois sur eux-mêmes et sur les autres membres de leur famille. Ces choix ont généralement trait à leur histoire démographique (mariage, divorce, fécondité), à leur activité (offre de travail, préparation de la retraite), à leur lieu de résidence (ville, quartier, choix entre maison et appartement, entre location et achat), ou encore à leurs investissements financiers ou immobiliers, à leurs héritages et autres donations, et à l'éducation de leurs enfants.

Ces choix effectués par les individus au cours de leur cycle de vie sont soumis à de nombreux aléas tels que la maladie et même le décès d'un membre du ménage, la perte d'emploi, la réussite scolaire de leurs enfants, les variations de la situation macroéconomique déterminant les prix (des biens de consommation et des logements), les salaires, les perspectives d'emploi, la rentabilité de leurs investissements financiers et immobiliers. D'autre part, les choix effectués par un membre du ménage ont souvent des conséquences importantes sur les autres membres du ménage, ce qui implique de considérer les décisions jointes au sein du ménage.

Afin d'étudier ces différentes questions, l'économiste doit isoler un objet d'étude, c'est-à-dire un ensemble d'interactions qui lui semblent particulièrement importantes.

Citons trois thématiques potentielles :

2.1.1. Localisation du ménage, investissements immobiliers, et offre de travail

Il s'agit ici de choix de moyen terme effectués au sein de couples soumis à un environnement incertain. Le risque et plus généralement l'incertitude portent à la fois sur l'emploi de chaque conjoint (localisation, horaires de travail et niveau de rémunération), sur l'évolution des prix de l'immobilier, et sur la conjoncture économique qui détermine les perspectives d'emploi.

L'augmentation du prix du pétrole constitue un sujet d'actualité susceptible de faire des trajets domicile-travail un élément de plus en plus important dans les choix de localisation. Lorsque l'on est locataire et célibataire, il est relativement aisé de déménager pour s'adapter à chaque changement d'emploi (localisation ou niveau de rémunération). Toutefois, les coûts de transaction (financière ou psychologique) liés à un déménagement, et les difficultés à trouver un logement dans certaines villes ou quartiers rendent moins optimal le déménagement systématique en réaction à chaque changement professionnel. Par ailleurs, la résidence principale peut être achetée plutôt que louée, ce qui lie les choix de localisation et d'investissement.

Dans ce cas, les coûts de transaction sont tels qu'il est difficile d'envisager de déménager à chaque changement professionnel. Les investissements immobiliers doivent alors s'effectuer dans une optique dynamique anticipant les changements professionnels, qu'ils soient choisis ou subis.

Tous les ingrédients sont ainsi réunis pour élaborer un programme (individuel) dynamique modélisant les interactions en matière d'emploi, d'investissement immobilier et financier et de localisation résidentielle.

Le problème se complique dans le cas des couples car il s'agit d'optimiser les trajets des deux conjoints, en tenant compte des interactions entre les offres de travail des deux conjoints, ainsi que des investissements immobiliers (et, par conséquent, des investissements financiers, qui sont plus ou moins substituables aux investissements immobiliers). Il s'agit alors d'élaborer un modèle permettant d'analyser la dynamique des choix joints du couple en matière d'emploi, d'investissements financiers et immobiliers, et de localisation résidentielle.

Les implications de ce modèle offriront un éclairage nouveau sur l'estimation des modèles de localisation résidentielle et d'arbitrage entre location et achat. En particulier, si la littérature a montré l'importance de l'accessibilité aux emplois dans les choix de localisation (voir de Palma et al., 2005 et 2007[s1] pour le cas de la région parisienne), elle s'est limitée soit à une définition très générale et non individualisée de l'accessibilité, qui ne peut pas prendre en compte les différences entre l'homme et la femme, soit à un simple temps de trajet entre le domicile et l'emploi actuel, qui ne prend pas en compte la dynamique des choix.

Le modèle théorique que nous souhaitons développer doit conduire à une définition individualisée de l'accessibilité bien distincte pour l'homme et la femme et cohérente avec les opportunités d'emploi de chacun, qui dépendent de son sexe, son âge, ses diplômes et son histoire professionnelle. Les données de recensement et d'enquêtes classiques telles que l'enquête Logement et l'enquête Emploi pourront être mobilisées efficacement pour estimer ce modèle à court terme, mais nous ne connaissons pas de sources de données françaises permettant de constituer un panel suffisamment riche (tel que, par exemple, le PSID américain) pour estimer la dynamique des choix.

Par ailleurs, des travaux ont montré que les implications et les conséquences professionnelles des mobilités différaient fortement entre célibataires et personnes en couples et au sein des couples entre les hommes et femmes [20]. Dés lors, raisonner au niveau du couple est essentiel pour comprendre les mécanismes et depuis les travaux pionniers de Mincer [18], les choix migratoires sont désormais

de plus en plus appréhendés au niveau du couple. La plupart des travaux empiriques portent sur des données de panels américains de grande ampleur.

2.1.2. Investissements financiers, transferts financiers et divorce

Nous nous intéressons ici aux différents transferts financiers au sein des ménages ou des familles, mais aussi entre les ménages, l'Etat ou les institutions financières. Ces transferts, plus ou moins volontaires, peuvent prendre la forme d'héritage, de donation entre vivants, de pension alimentaire, d'assurance-vie, de cotisation retraite (dans le cas de pension de réversion) ou d'autres investissements financiers ou immobiliers.

L'héritage, l'assurance-vie, les donations entre vivants et autres pensions peuvent être vus comme autant de moyens plus ou moins substituables de transfert entre les différents membres de la famille. Dans cette optique, toute modification de législation (taux de prélèvement de l'Etat, avantages fiscaux, règles de calcul des pensions alimentaires, désignation des héritiers ou bénéficiaires par défaut, niveau de protection face aux risques, etc.) concernant l'un de ces vecteurs de transferts est susceptible d'influencer les choix individuels concernant l'ensemble des transferts. En disposant de panels longs, il est possible d'exploiter le fait que la législation touche différemment les différents individus au cours de leur cycle de vie et en fonction de leur situation économique et démographique (composition de leur famille nucléaire et plus large) pour estimer des modèles de « différence en différence » ou construire des scénarios d'expériences quasi naturelles.

L'accès à des données de panels longs dans ces domaines permettrait d'analyser la dynamique des choix en matière de transferts intra-familiaux, et de comprendre l'influence de la législation sur le divorce ou l'héritage sur l'ensemble des transferts intrafamiliaux.

Les panels actuellement disponibles en France comme les panels européens ECHP puis SILC ont des effectifs souvent insuffisants pour traiter des événements assez rares comme la mobilité de longue distance ou la séparation du couple. Par exemple, dans l'ECHP, seulement 60 couples ont connu un divorce au cours de la période 1994-2001 [24].

2.1.3. Mariage, fécondité et travail

L'existence d'interactions importantes entre le travail, le mariage, et la fécondité est généralement reconnue par les économistes, même si ces interactions sont souvent mal comprises. Un nombre considérable d'études empiriques montrent effectivement que les hommes mariés travaillent un plus grand nombre d'heures, et les femmes mariées un plus petit nombre, que leurs correspondants célibataires. Par ailleurs, des modèles théoriques (à la suite des travaux précurseurs de Becker, 1973) montrent que la décision d'une personne de se marier est une fonction de sa position sur le marché du mariage et donc de son salaire. Quant au choix d'avoir des enfants, il dépend naturellement du statut marital mais également des perspectives sur le marché du travail de l'un et l'autre des conjoints. L'analyse de ces interactions dynamiques peut aider, par exemple, à mieux comprendre les causes de la diminution de la fécondité observée ces dernières années en France comme dans la plupart des pays occidentaux, et d'identifier les politiques économiques les mieux adaptées. De plus, le déclin de la fécondité peut s'expliquer par le risque de divorce. Cependant, ce risque est également endogène car, par exemple, les conjoints peuvent le réduire en choisissant d'avoir plus d'enfants.

Dans le même ordre d'idées, on peut penser que les femmes développent en travaillant une assurance contre une décision unilatérale de divorce de leur partenaire ou que, inversement, la l'amélioration de la situation financière des femmes (par augmentation de leur offre de travail) participe à l'accroissement du taux de divorce. Ces différentes idées peuvent être testées sur la base d'estimations sur données de panel telles le PSID ou le NLSY. De plus, si l'on dispose de données sur une période suffisamment longue, il sera possible d'utiliser les changements dans la législation sur le divorce comme variation exogène du risque de divorce.

Le PSID français ne semble pas pour demain ; l'idée de tout savoir sur un échantillon a été le rêve d'anthropologues ; telle fut l'opération Plozevet [15]. A-t-elle été un tel succès ? Ce fut à nouveau l'idée sans suite effective de prolonger jusqu'à nos jours la vaste enquête de démographie historique TRA de Jacques Dupâquier relative au XIX^{ème} siècle.

3. Suppléer au manque de panels par des micro-simulations

Quand la collecte ne peut prendre l'extension rêvée, il faut se résoudre à mettre en place une vision simplifiée sous certaines hypothèses, c'est à dire sous un modèle. C'est l'histoire des sondages en termes d'extension démographique ; c'est celle plus récente des modèles de micro-simulation en termes de champ thématique couvert ; on parle aussi de fusions de données ou d'enquêtes [16] ; il s'agit de prêter à des individus des comportements observés dans d'autres enquêtes sur d'autres personnes comportant un nombre suffisant de variables communes servant de régresseurs. Ces modèles servent aux projections, tant la statistique est faite pour prédire les déplacements ou la survie des populations [3], l'avenir des retraites, mais aussi les comportements de consommation ou d'offre de travail [11], [12], [13], [14].

Ils peuvent pallier le manque de données dans certains cas. Ils sont un bon outil pour estimer par exemple les transferts publics et simuler des politiques économiques mais restent limités pour étudier les interactions entre différents domaines. Ils représentent plutôt un complément qu'un substitut à l'appariement.

Conclusion

La lecture de la lettre du CNIS Chroniques n° 8 fait ressortir l'insatisfaction des statisticiens à ne pouvoir joindre les avantages respectifs des deux familles de sources des enquêtes sur l'emploi et des fichiers administratifs relatifs à l'emploi et au chômage. Nous proposons une solution fondée sur des appariements sécurisés permettant à la fois d'évaluer au niveau individuel les divergences entre sources, d'enrichir les fichiers administratifs des concepts et des informations de l'enquête emploi sur un échantillon national et de les projeter au niveau local grâce aux techniques des estimations sur petits domaines diffusées en France par Pascal Ardilly.

On pourrait envisager de continuer à abonder le fichier de l'enquête emploi, même une fois la collecte par questionnaire arrêtée avec les mêmes données administratives. On créerait ainsi une source longitudinale très riche sur l'emploi. La question de l'évolution de la composition du ménage reste ouverte, à moins de recourir à la source fiscale.

Certes, cette source ne satisferait pas toutes leurs demandes, mais les modèles de fusion d'enquêtes et de micro-simulation leur ouvre – bien sûr sous modèle et leurs hypothèses simplificatrices d'indépendance- des méthodes pour aller vers les résultats cherchés.

Bien entendu, la diffusion de telles sources doit être justifiée et surveillée. Elle peut ou doit être sécurisée dans un centre spécialisé. Mais si sa constitution devient possible, cela permettra à de nombreux chercheurs de travailler enfin sur données françaises.

Bibliographie

[1] Ardilly P., « Panorama des principales méthodes d'estimation sur les petits domaines », INSEE, Méthodologie statistique, n° M0602, cours du CEPE, septembre 2006.

[2] Becker 1973

[3] Bonneuil N., « Jeux, équilibres, et régulation des populations sous contraintes de viabilité : une lecture de l'oeuvre de l'anthropologue Fredrik Barth », Population, revue de l'INED, n° 4, pp 947-976, juillet-août 1997.

[4] Chaleix M., Lollivier S., « Des panels pour les statistiques sociales », Le Courrier des statistiques, n° 113-114, pp 53-56, mars-juin 2005.

[5] CNIS, « Emploi, Chômage, précarité. Mieux mesurer pour mieux débattre et mieux agir. Présentation du rapport du groupe présidé par Jean-Baptiste de Foucauld », Chroniques n° 8, pp 1-6, octobre 2008 et Rapport du CNIS, n° 108, septembre 2008.

[6] CNIS, « Transcription en droit français de la directive européenne n°95/46/CE du 24 octobre 1995 », Rapport du CNIS n° 55, janvier 2000.

[7] Dupâquier J., « La généalogie au service de l'histoire » in « La généalogie : une passion française » / dir. Marie-Odile Mergnac. – Paris, Autrement, p. 101-112, 2003.

[8] Gensbittel M-H., Riandey B., Quantin C., « Appariements sécurisés : statisticiens ayez de l'audace ! », Le Courrier des statistiques, n° 121-122, pp 49-58, mai-décembre 2007.
[9] Goldberg M., Quantin C., Guéguen A., Zins M., "Bases de données médico-administratives et épidémiologie : intérêts et limites », Le Courrier des statistiques, n° 124, pp 59-70, mai-octobre 2008.
[10] Goy A., « L'appariement sécurisé de fichiers d'étudiants grâce au hachage des identifiants », Le Courrier des statistiques, n° 113-114, pp 23-32, mars-juin 2005 et Journal de la Société française de statistique, vol 146 n°3, 2005.
[11] Gravel, N, C. Hagneré, N. Picard (2004), "Une estimation des conséquences d'une réforme des minima sociaux sur l'offre de travail à l'aide d'un modèle intertemporel de microsimulation", *Economie Publique, Etudes et recherches*, 2004/1, p. 3-44.
[12] Gravel, N, C. Hagneré, N. Picard, A. Trannoy (2001) " Une évaluation de l'impact incitatif et redistributif d'une réforme des minima sociaux ", *Revue française d'économie,* vol. XVI, pp. 125-167.
[13] Hagneré C. (2001), "SIMPTOM : un modèle de microsimulation avec information temporelle", mimeo.
[14] Hagneré C., Picard N., Trannoy A., Van Der Straeten K. (2002) "La prime pour l'emploi est–elle optimale ?", *Economie Publique, Etudes et Recherches*, 11.
[15] Kourganoff M. et JC., « Rapport d'observation psychologique effectué dans la commune de Plozévet (Sud Finistère), rapport de l'Ined , 347 p, 1963.
[16] Lejeune M. (dir), « Traitements de fichiers d'enquêtes. Redressements, injections de réponses, fusions », Presses universitaires de Grenoble, Collection Libres cours, 128p, 2001.
[17] Lenormand F., « Le système d'information de l'assurance maladie », Le Courrier des statistiques, n° 113-114, pp 33-52, mars-juin 2005.
[18] Mincer J. 1978. Family Migration Decision. *The Journal of Political Economy*. 86(5): 749-773.
[19] Mizrahi A. et A., « Premiers sondages français dans les dossiers de sécurité sociale et appariement avec les enquêtes auprès des ménages », in Lavallée P. et Rivest L.P. (dir) « Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine » Actes du 4ème colloque francophone sur les sondages de la Société Française de Statistique, pp 117-127, Dunod, Paris, 2006.
[20] Pailhé A., Solaz A. 2008, "Professional Outcomes of Internal Migration by Couples: Evidence from France", *Population, Space and Place*, 14:347-363.
[21] Quantin C., « Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales », VIIème Journées de Méthodologie Statistique de l'INSEE, session 3, décembre 2000 ou Courrier des statistiques n°113-114, juin 2005 et Journal de la Société française de statistique, vol 146 n°3, 2005.
[22] Quantin C., « Linking multiple and heterogeneous databases : confidentiality and patient identification issues ». Open University "Statistical methods for health registers and linked databases ", 20-21 mai 2009.
[23] Quantin C., Fassa M.,. Coatrieux G., Riandey B., Trouessin G., Allaert F.A., « Chainage de bases de données anonymisées pour les études épidémiologiques multicentriques nationales et internationales : proposition d'un algorithme cryptographique », Revue d'épidémiologie et de santé publique, vol 57, pp 33-39, 2009.
[24] Uunk W. (2004), "The economic consequences of Divorce for Women in the European Union: the Impact of Welfare State Arrangements", *European Journal of Population* 20: 251-285