Cahiers de démographie locale 2008, Néothèque, coll.
« Dynamiques des populations locales », 2008, pp. 205-220.

Méthode sécurisée de traitement des fichiers administratifs

Benoît RIANDEY*

Dans les articles précédents, les démographes locaux ont cité des applications mobilisant plusieurs sources administratives. Il s'agit de dénombrer à l'adresse et par âge les enfants à scolariser dans la commune ou de repérer les migrations internes à l'agglomération. Ce peut être aussi le projet de rassembler des informations relatives aux mêmes personnes ou ménages extraites de fichiers administratifs distincts. Compter sans double compte, donc en éliminer les doublons ; repérer l'identité commune à plusieurs enregistrements pour fusionner leurs informations, opérer un chaînage longitudinal. Ce sont bien les opérations de base de la démographie. Réalisées dans une institution à partir de ses propres fichiers, elles ne posent que les difficultés ordinaires de la statistique. L'identité de la personne ou un identifiant unique dans l'institution permet de fusionner les données, repérer les enregistrements venant en rebut et remédier aux erreurs d'identification.

Les difficultés sont tout autres lorsque les sources sont réparties entre institutions. D'une part, les éléments d'identification ne sont plus communs. Bien plus, l'accès aux données personnelles soulève d'importants problèmes juridiques.

On objectera que le 6 août 2004, la transcription de la directive européenne du 24 octobre 1995 a reconnu la légitimité statistique d'accéder à des informations nominatives extérieures. Dans son article 6, la loi Informatique et Libertés modifiée stipule bien que : « Un traitement ultérieur de données (à caractère personnel) à des fins statistiques ou à des fins de recherche scientifique ou historique est considéré comme compatible avec les finalités initiales de la collecte des données, s'il est réalisé dans le respect des principes et des procédures ». Les démographes locaux peuvent bien sûr se prévaloir de

* Institut national des études démographiques

cet article pour demander accès aux données nominatives des institutions concernées, puis engager les démarches adéquates auprès de la CNIL.

Le succès pourrait ne pas être grand. Bien qu'autorisées par la loi, les administrations ont toutes chances d'être réticentes à transmettre les données à caractère personnel de leurs administrés. D'ailleurs, si elles y sont tenues de par l'ordonnance du 25 mars 2004 modifiant la loi de statistique publique de juin 1951, c'est uniquement à l'égard de l'INSEE et des services statistiques ministériels. Les instituts de recherche et les services statistiques des collectivités locales ne peuvent prétendre à bénéficier de cette prérogative de l'obligation statistique, seulement du bon vouloir.

Obtenir des traitements faits chez elles par ces institutions se montrerait inopérant pour l'objectif d'appariement annoncé. Obtenir un extrait anonyme des fichiers pourrait être négociable, mais serait aussi vain pour cet objectif.

Eh bien non justement. On peut trouver des solutions pour réaliser des appariements entre fichiers anonymes. En premier abord, on pensera qu'en présence de fichiers exhaustifs très riches en informations identifiantes, comme par exemple la date de naissance détaillée, on peut inférer l'identité de la personne. Notre propos n'est pas de mettre en œuvre cette démarche frauduleuse de l'identification indirecte sauvage. Il s'agit pour nous, grâce à un identifiant secret inviolable, de reconnaître légitimement que deux informations sont relatives à la même personne sans qu'on puisse inférer de qui il s'agit. Les épidémiologistes ont été conduits à mettre au point ces méthodes sur des données extrêmement sensibles et relevant d'une très haute priorité pour la politique de santé publique.

L'expérience des épidémiologistes

Le développement de l'épidémie de sida a exigé une veille épidémique assortie des plus grandes précautions de confidentialité en raison du caractère stigmatisant de cette maladie. C'était particulièrement le cas en région parisienne où le foyer épidémique concernait en priorité la communauté homosexuelle masculine et dans le Midi où c'étaient les toxicomanes pratiquant le partage des seringues. Dès l'année 1988,

l'équipe marseillaise du docteur Thirion proposait un algorithme donnant aux personnes concernées un identifiant déduit du patronyme sans qu'on puisse retourner au nom. Ainsi une personne qui aurait été diagnostiquée positive au VIH sida pour plusieurs prises de sang aurait un même identifiant dans les multiples enregistrements anonymes de ses tests sanguins et apparaîtrait une seule fois comme un nouveau cas de l'épidémie. Inversement dans un registre du cancer, les examens successifs d'un même patient pourraient être réunis dans le même dossier en l'absence du nom ou du numéro INSEE, le Nir.

L'algorithme marseillais a été rapidement repris au niveau national par Jacques Valleron, mais il ne s'est pas montré parfait : la fonction de compression de l'identifiant qui préserve de toute fonction inverse de retour au patronyme était excessive et provoquait des « collisions » ; c'est-à-dire qu'elle permettait que plusieurs sujets puissent se voir attribuer le même identifiant, même en partant d'une information nominative très redondante avant ce cryptage. Une précaution de la procédure marseillaise consistait d'abord à neutraliser les fautes orthographiques : le patronyme, enregistré de manière variable, aurait abouti à des identifiants distincts inconciliables, donc à des doublons d'enregistrements pour la même personne. Aussi, avant cryptage du patronyme, procédait-on à sa réduction phonétique. Certes, cette excellente précaution risquait de générer des doublons, ce qui a rendu d'autant plus utile la précaution de partir d'une information suffisamment redondante et la recherche d'un meilleur algorithme d'anonymisation.

L'évolution du contexte

Cette première expérience a bénéficié de deux développements favorables ; en premier lieu, la libéralisation du cryptage par l'Etat français est intervenue en 1998-1999. Egalement, des algorithmes plus sûrs ont été développés, en particulier celui de Shamir préconisé par Jaro (1995) pour les fichiers épidémiologiques. Le risque de collision est devenu atomique, de l'ordre de 10^{-48}.

Ces nouveaux algorithmes SHA constituent maintenant le moteur des appariements sécurisés développés plus récemment par l'assurance

maladie (Trouessin, 1997, Lenormand, 2005) et par le CHU de Dijon (Quantin, 1998, 2005).

Pour l'assurance maladie, il s'agissait de rassembler anonymement les informations relatives à tous les séjours hospitaliers d'une même personne dans un quelconque établissement privé ou public. C'est le programme du PMSI(1).

Au CHU de Dijon, Catherine Quantin, médecin responsable du département de l'information médicale, et son équipe se sont mobilisés pour une dynamique régionale, la mise en cohérence des informations individuelles relatives à la périnatalité d'un réseau de 18 établissements publics ou privés en Bourgogne (Gouyon, 1999). Enfin, il s'agissait de dénombrer la file active en cancérologie d'un hôpital ou l'inter-file active d'un réseau d'hôpitaux. A nouveau l'objectif cherché visait à dénombrer sans double compte de façon anonyme des personnes présentes dans plusieurs institutions.

L'assurance maladie et le CHU de Dijon, tout en utilisant le même algorithme SHA ont eu recours à deux méthodologies bien différentes qui éclairent les choix s'offrant aux démographes locaux.

Deux méthodes d'appariement sécurisé

On distingue deux méthodes d'appariements, *l'appariement déterministe* à l'image de la fonction FOIN et *l'appariement probabiliste* illustré par la méthode ANONYMAT(2).

La fonction FOIN(3) se fonde sur un identifiant composé du numéro Nir de l'assuré complété de la date de naissance complète du patient et de son sexe ; puis cette chaîne de caractères est hachée en un seul bloc par l'algorithme mathématique SHA (Jaro, 1995). La même chaîne de caractères hachée en utilisant la même clé de hachage conduit toujours au même identifiant FOIN. Au contraire, un utilisateur qui ne disposerait pas de la même clé aboutirait avec la même chaîne d'identification

(1) Programme médicalisé du système d'information (hospitalisation publique ou privée).

(2) La présentation qui suit des méthodes de hachage est reprise de ma communication commune avec Catherine Quantin à la Chaire Quetelet 2006. C'est son application aux données locales qui constitue l'apport de ce texte.

(3) Fonction d'occultation des informations nominatives (Trouessin, 1997, Tardif, 2007).

initiale à un autre identifiant FOIN. Ce procédé fournit donc à la fois la fidélité dans l'identification, la protection par rapport aux usages non prédéfinis par le partage de la même clé et l'anonymat définitif.

L'identifiant FOIN permet de rassembler dans le même fichier statistique du PMSI tous les épisodes d'hospitalisation qu'auraient connus le même patient. De même, grâce à cet identifiant FOIN et malgré la pléthore de régimes et d'institutions, l'assurance maladie a pu mettre en place une base de données unifiée, exhaustive et anonyme des consommations médicales triennales des 60 millions de bénéficiaires, le SNIIR-AM(4) (Lenormand, 2005). Bien sûr, il s'agit d'un entrepôt de données herculéen, qui bientôt accueillera le PMSI et dont les applications se mettent en place progressivement, stimulées par le nouvel IDS, l'Institut des données de santé.

La puissance exhaustive du SNIIR-AM n'en fait pas un outil manipulable pour les aller-retours incessants de la recherche. Le sondage s'impose donc. C'est l'EPIBAM(5), un panel au 100ème du SNIIR-AM alimenté non plus sur trois années de consommation médicale mais à terme sur vingt ans. (Lenormand, 2005).

Ce dispositif statistique impressionne, mais il comprend des limites : l'EPIBAM constitue un vaste panel et une base de sondage adaptée aux enquêtes en population générale. Mais elle est trop restreinte pour les petites sous-populations. Dans l'EPIBAM, les maladies orphelines demeureront orphelines de sondage. Seul le dossier médical personnel pourra dépasser cette insuffisance grâce à sa diffusion exhaustive.

Le SNIIR-AM est un colosse encore immature : l'identifiant du bénéficiaire varie avec son statut : d'abord ayant droit, celui-ci devient un jour lui-même assuré : c'est le cas du conjoint au foyer et des enfants. Cette situation est aggravée pour les enfants de divorcés qui, avant de s'émanciper, naviguent entre les deux Nir parentaux.

De ces remarques, on tire trois conclusions :

(4) Système national inter-régime d'information de l'assurance maladie

(5) Echantillon permanent inter-régime des bénéficiaires de l'assurance maladie. L'unité statistique retenue est l'individu. Ce choix le distingue de façon essentielle de l'EPAS (Mizrahi, 2006), l'Echantillon permanent des assurés sociaux, qui repose sur l'assuré, une unité instable, intermédiaire entre le ménage et l'individu et donc peu propice à l'analyse démographique.

- l'identifiant doit être stable. Prochainement, le « SNIIR-AM 2 » s'appuiera sur le Nir du bénéficiaire après diffusion générale d'une carte « Vitale 2 » de bénéficiaire (et non plus d'assuré), toujours fondée sur le Nir personnel. Ce sera la maturité ;
- on doit prévoir une procédure de retour des enregistrements en échec d'appariement auprès des institutions productrices des sources nominatives ;
- un identifiant très redondant multi-modulaire faciliterait la correction des erreurs.

Cette dernière conclusion est à l'origine de la seconde méthode d'appariement sécurisé, mise en œuvre au CHU de Dijon avec le même algorithme SHA. C'est la procédure Anonymat. L'appariement déterministe FOIN exige la concordance exacte de la chaîne complète d'identifiants. Au contraire, la procédure Anonymat hache séparément les différents modules de l'identifiant, qu'elle pondère en fonction de la valeur discriminante de chacun d'entre eux. Pour chaque paire d'enregistrements, elle évalue alors, en fonction du nombre d'éléments concordants, la probabilité d'un appariement exact entre ces deux enregistrements, puis répartit les paires d'enregistrements en trois lots : appariés, certainement distincts, incertains. Cette procédure fine est donc recommandée lorsque les données du fichier sont incomplètes ou entachées d'erreurs.

Les modules composant l'identifiant sont assez souvent les suivants : Nom, Prénom, Date de naissance. Le sexe, souvent erroné, compact seulement avant cryptage, est écarté par l'équipe de Dijon pour son très faible pouvoir discriminant : on est moins souvent homonyme que de même sexe ! Pour ce type d'identifiant s'ajoute le risque des erreurs d'orthographe : une lettre fausse et les identifiants cryptés deviennent méconnaissables. Comme pour le sida à Marseille, on procède donc d'abord à un traitement phonétique qui simplifie les orthographes, au risque de quelques homonymes que la redondance de l'identifiant permet de distinguer.

L'Institut de veille sanitaire a recours à ces techniques pour l'enregistrement des maladies à déclaration obligatoire. L'identifiant, composé de l'initiale du nom, du prénom, de la date de naissance et du sexe, est haché par Anonymat.

L'identifiant à composante familiale (Quantin, Gouyon, 2006) est une forme complexe d'identifiant modulaire ainsi composé : nom de naissance d'ego, premier prénom d'état civil, date de naissance, puis nom, prénom, dates de naissance du père, puis de la mère. Cette extension repose d'abord sur l'expérience pédiatrique qui conduit à associer l'identifiant de l'enfant à celui de sa mère, l'un et l'autre pouvant faire l'objet de traitements spécifiques. Pour les études génétiques, ajouter l'identifiant du père s'impose. Cet identifiant, faisant l'objet d'un brevet déposé par le CHU de Dijon et la start-up HC Forum d'Olivier Cohen, s'avère précieux pour les études internationales, notamment européennes, pour l'épidémiologie génétique et la démographie. Il est mobilisé pour l'Observatoire (longitudinal) de l'enfance en danger de l'ONED.

La carte de santé électronique européenne exigera un identifiant stable indépendamment du pays actuel d'assurance maladie de la personne. Les numéros nationaux sont donc inappropriés ; cet identifiant modulaire et anonymisé constitue un candidat numérique adéquat. Validé par la CNIL, il donne accès depuis l'étranger, avec l'accord du patient, à ses informations médicales sur la plate-forme Internet sécurisée HC Forum (Gensbittel, 2007). Cette plate-forme est destinée aux études européennes des maladies génétiques rares.

Le hachage de l'identifiant ne suffit pas à assurer la confidentialité totale : l'institution hacheuse est susceptible de détenir la table de correspondance entre identifiants bruts et hachés ; elle pourrait aussi procéder à une attaque par dictionnaire : s'intéressant à une personne, elle pourrait hacher son identifiant et repérer dans ses fichiers les enregistrements contenant cet identifiant haché. Pour cette raison, on pratique un double hachage amont avant transfert, aval après réception. Aucune institution ne détient alors les éléments d'identification. On a ainsi créé l'anonymat à l'égard de tous. Enfin le croisement de certaines informations anonymes précises du fichier peut permettre une identification indirecte, mais ceci n'est pas spécifique aux données appariées qui doivent à cet égard faire l'objet des mêmes précautions que les autres fichiers.

L'état de ces avancées est rapporté par la double publication du *Courrier des statistiques* et du *Journal de la Société Française de Statistique* en

2005. Ce point assez complet étant fait, revenons à la démographie locale.

Application à la démographie locale

A la fin des années 90, le CREDOC a été chargé de créer un observatoire du RMI à Paris. A cet effet, il devait rassembler des données individuelles relatives aux allocataires parisiens du RMI, le revenu minimum d'insertion. Ces informations nominatives sont réparties entre de multiples institutions. Leur appariement aurait donc soulevé un problème difficile en droit français. C'est la CNIL (rapport, 2001) qui a suggéré d'éviter cet obstacle juridique en appariant des données anonymisées et, à cet effet, de recourir aux méthodes mises au point par le CHU de Dijon (Aldeghi, 2004) : ces méthodes d'anonymisation, fécondes en épidémiologie, ne le sont, en effet, pas moins pour la statistique publique. L'INED, la Société française de statistique et le Département d'Information Médicale du CHU de Dijon ont diffusé cette idée simple au cours d'une série de séminaires, tables rondes, cours, présentation aux Journées de méthodologie statistique de l'INSEE.

Cet exemple du CREDOC est certainement reproductible dans bien des problématiques locales car les observatoires sont très fréquemment dédiés à des phénomènes régionaux ou locaux. D'ailleurs, mobiliser les fichiers administratifs est financièrement plus accessible à une petite collectivité que la réalisation d'une enquête.

Le ministère de l'Education nationale fut le premier service statistique ministériel à mettre en œuvre ces techniques (Goy, 2005). Elles s'avèrent précieuses pour articuler les niveaux national et local, qu'il s'agisse des rectorats ou des universités. D'une part, l'Education nationale a dû élaborer un identifiant des élèves ou étudiants spécifique à son secteur, l'Ine, mais sans que lui soit permis une libre circulation entre les échelons locaux et national des données ainsi identifiées : pour les élèves, il ne lui était pas possible de rassembler les éléments d'un cursus qui se serait déroulé à travers plusieurs rectorats. Pour les étudiants, le service statistique avait le droit de reconstituer les cursus des étudiants mobiles dans son fichier SISE, mais pas celui de les communiquer aux universités. Chaque université ignorait ainsi le parcours de ses étudiants à la sortie de cette université ; le taux d'échec apparent de ses étudiants

en est gravement surestimé. Ces carences ont été comblées par le hachage des Ine qui autorise la circulation de ces données complètes mais anonymisées. Les 99 universités françaises peuvent ainsi suivre statistiquement le devenir des étudiants de leurs diverses filières. Il s'agit donc d'un exemple intéressant soulignant la difficulté d'articuler des données longitudinales régionales et nationales ou tout simplement inter-régionales. Mais cet exemple montre qu'une solution a été trouvée grâce aux appariements sécurisés.

Par contre, cette application ne permet pas de repérer les doublons d'étudiants inscrits dans deux universités sous deux numéros Ine. La gestion d'un identifiant spécifique couvrant tout un secteur soulève des difficultés analogues à celles d'un recensement qui se doit d'être exhaustif et sans double compte. C'est bien la force du Nir que de certifier que deux valeurs distinctes de l'identifiant distinguent toujours deux personnes distinctes. C'est d'ailleurs pour cette vertu que le Nir a été réintroduit au ministère des Finances afin de dédoublonner l'identifiant fiscal.

A l'Education nationale, pour dédoublonner l'Ine, il aurait aussi fallu une procédure de validation externe de l'Ine fondée sur le Nir, ou seulement sur un Nir haché(6).

En fait, le service statistique du ministère a développé une stratégie plus ambitieuse : il a mis en œuvre en 2006-2007 une base exhaustive des identifiants Ine : identifiant national élève-étudiant. C'est donc un répertoire général de la population passée depuis 2006 dans le système éducatif français. L'objectif en est de stocker dans la future base FAERE l'ensemble des trajectoires éducatives complètes de tous ces jeunes ; c'est grâce à l'Ine haché que les morceaux de trajectoires individuelles peuvent être recollés. Ce projet initié par Alain Goy est considérable.

La construction d'un répertoire national n'est pas une mince affaire. Certes, l'entrée dans le système éducatif présente, comme la naissance, le caractère universel et unique propice à la création d'un répertoire, mais l'entrée ne se fait pas toujours en petite maternelle… notamment pour les étudiants étrangers. Et les étudiants peuvent s'inscrire

(6) En fait, la saisie auto-administrée du Nir, aussitôt haché, et celle de l'identifiant sectoriel Ine permettent l'élimination anonyme des doublons sans rapprochement du RNIPP car l'unicité du Nir est durablement validée, mais cette voie innovante et générale de validation des identifiants n'a, semble-t-il, encore jamais été employée.

simultanément dans plusieurs établissements. La consultation d'un répertoire est alors une bonne solution pour éviter toute identification multiple des étudiants déjà connus de l'enseignement français. Pour les étudiants venant de l'étranger, cette immatriculation appelle une méthodologie rigoureuse.

Les fichiers de scolarité des écoles utiliseront donc cet identifiant des élèves. Des extractions, munies de l'identifiant haché avec une clé spécifique, permettraient aux collectivités locales de compter les élèves présents sur leur territoire. Son apport à la démographie locale mérite examen, mais comprenons bien que, comme le RNIPP et le RNIAM, ce n'est pas un registre de cette population, dans le sens qu'il ne comprend pas l'adresse des élèves.

En définitive, cet exemple nous montre que, s'il n'est pas fondé sur un répertoire, l'identifiant ne présente pas les qualités censitaires. Ce constat justifie le recours à l'appariement probabiliste et aux identifiants modulaires utilisant le plus souvent le patronyme.

Illustrons ce constat par une belle opportunité de mesure des migrations : les appariements sécurisés offrent une solution élégante et simple pour un territoire accessible seulement par avion, c'est-à-dire une île isolée. C'est le cas du département de la Réunion. Imaginons qu'à l'aéroport de la Réunion, à l'arrivée comme au départ, le passager saisisse son identification, aussitôt hachée comme un mot de passe ou, que cette information anonyme soit produite à partir des enregistrements liées au déplacement. Cet identifiant (identique à l'arrivée et au départ, mais anonyme) pourvu des dates d'arrivée et de départ permet d'apparier une entrée et une sortie d'une même personne (dans l'ordre quelconque). On distingue alors par durée de présence, d'absence ou de séjour, les séjours en cours à la Réunion, les « émigrations » en cours hors de la Réunion et tous les séjours achevés, notamment touristiques. L'anonymat de la procédure permet ce rapprochement des « volets » d'entrée et de sortie sans entrer dans une procédure policière. Elle se fonde sur le hachage du nom et du prénom. Pourrait-on ajouter le sexe, l'âge de la personne, le fait d'être ou non né à la Réunion ? Cela demande réflexion, mais même sans ces compléments, l'information démographique produite si aisément serait de première importance. Il faut toutefois être attentif à la qualité de la saisie du nom et du prénom. Une variation entraverait l'appariement et transformerait

par exemple un séjour touristique en une immigration durable et une migration durable en un séjour touristique. Il ne faudrait pas que quelques erreurs sur un gros volume de voyages perturbent la mesure du phénomène étudié, de bien moindre fréquence, les migrations.

A l'heure actuelle la mesure classique des flux d'entrées et de sorties permet seulement de calculer un solde qu'on ne peut même pas appeler migratoire car il ne distingue pas les séjours de courte durée, notamment touristiques, des migrations (séjours d'au moins un an). La mise au point de cette procédure constituerait un progrès remarquable.

Stratégie à développer

Le choix entre les deux méthodologies de hachage dépend donc d'abord des variables d'identification disponibles et de leur qualité. La disponibilité du Nir dans le circuit hospitalier a conduit au choix de FOIN, identifiant haché en un bloc. Mais ce n'est pas sans problème, comme en fait part Tardif (2007) : l'usage du Nir ne va pas sans erreur et près de 5 % des séjours hospitaliers de l'exercice 2004 n'ont pu être chaînés, mais le système est en amélioration constante. En témoigne également l'étude de Jean-Paul Beyeme-Ondua sur l'évaluation de la qualité des données chaînées du cancer colorectal. Les séjours de l'année 2003 non chaînés de ce champ limité s'élevaient à 6,5 %. Ces échecs d'appariements sont la contrepartie du choix de couvrir l'ensemble du pays grâce au PMSI, alors que les registres de cancers, plus rigoureux, ne concernent que quelques départements, soient 15 % de la population.

Evoquons une situation que pourraient rencontrer les collectivités locales : l'ANAEM (ex-Office des Migrations Internationales) souhaitait comparer ses effectifs d'immigrants avec ceux du fichier AGDREFF des titres de séjour du ministère de l'Intérieur. Il s'agit en principe de la même population et les effectifs coïncident, mais sans qu'on soit sûr d'avoir compté les mêmes personnes. L'identifiant est commun, mais on veut procéder de façon anonyme pour éviter toute complication juridique : il suffirait de hacher cet identifiant avec la même clé, puis d'apparier les deux listes d'identifiants hachés pour compter l'effectif communs aux deux fichiers. Si on veut qualifier les cas non appariés, on fusionne deux extraits de fichiers munis de l'identifiant haché. N'est-il pas fréquent que des institutions locales soient prêtes à une

collaboration statistique sans se sentir le droit de dévoiler leurs fichiers ? Il y a donc pour elles moyen de s'entendre. Par contre, ces institutions n'ont le plus souvent pas d'identifiant commun. L'exemple du réseau de périnatalité de Bourgogne illustre alors la marche plus complexe à suivre.

Pour cette application, l'équipe du CHU de Dijon ne disposait pas du Nir ni de tout autre identifiant. Elle a donc construit un identifiant pluri-modulaire comprenant d'abord le nom, puis elle a eu recours à un chaînage probabiliste. C'était également la situation rencontrée par Isabelle Fournel : elle devait récupérer auprès de l'INSEE le statut vital et les causes de décès d'une cohorte de malades du cancer, cinq ans après le diagnostic ; en l'absence du lieu de naissance des patients et donc de leurs informations complètes d'état civil, elle ne pouvait recourir au circuit homologué passant par le répertoire de l'état civil puis par le service CepiDc de l'INSERM. Elle a donc appliqué la méthode Anonymat de Catherine Quantin en appariant son fichier de patients aux cinq fichiers annuels de mortalité de l'INSEE. Elle a obtenu un excellent taux de succès puisque son traitement automatique a bien classé 98,3 % de ses 10 000 patients et qu'après rattrapage manuel des cas indécis, elle a atteint un taux de succès de 99,2 %. Ceux-ci ont nécessité un retour aux dossiers des services de l'hôpital pour compléter ou corriger l'information, puis les enregistrements concernés ont fait l'objet d'un nouveau hachage et d'une nouvelle tentative d'appariement.

Rappelons qu'à l'Institut de veille sanitaire, on applique maintenant la méthode Anonymat pour le dénombrement des nouveaux cas de maladies à déclaration obligatoire ; mais on ne recueille que les initiales du nom et prénom, la date de naissance et le sexe du patient. C'est une information beaucoup plus facile à demander. On évite d'avoir à phonétiser le nom et le prénom pour neutraliser les fautes d'orthographe ; cependant on doit toujours être attentif au risque de nommer les femmes tantôt par leur nom d'état civil, tantôt par leur nom marital ; de même, le choix du prénom énoncé peut varier, sans même évoquer les diminutifs.

On doit évidemment être attentif au fondement même de cette technique du hachage irréversible. Elle n'admet pas de retour en arrière : le flot d'appariements s'effectue, en effet, toujours en aval à partir de fichiers antérieurement choisis en amont et hachés avec des clés

d'appariement compatibles. Par exemple, on ne peut pas sélectionner des individus du fichier haché pour retourner auprès d'eux collecter de nouvelles informations. Or dans certaines situations, on peut donc être conduit à « remonter le courant », c'est-à-dire à rompre l'anonymat. C'est notamment le cas des enquêtes épidémiologiques lorsqu'il apparaît qu'un patient requiert des soins pour un diagnostic porté au cours du protocole d'enquête. Il faut alors recourir à des tiers de confiance comme dans le projet de la grande cohorte d'enfants Elfe. En définitive, au delà des algorithmes à mettre en œuvre, c'est un dispositif à élaborer avec des personnes expérimentées. Elles existent comme en témoigne notre article du *Courrier des statistiques* en 2007 auquel nous avons également largement emprunté nos exemples (Gensbittel, 2007).

Références bibliographiques

Dossier spécial « Panels et appariements sécurisés », *Courrier des statistiques* n°113-114, juin 2005 (www.insee.fr) et *Journal de la Société française de statistique*, vol 146 n°3, 2005, coordonné par Benoît RIANDEY :

- QUANTIN Catherine, GOUYON Béatrice, ALLAERT François-André, COHEN Olivier, « Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales ».
- GOY Alain, « L'appariement sécurisé des fichiers d'étudiants grâce au hachage des identifiants ».
- LENORMAND François, « Le système d'information de l'assurance maladie ».

ALDEGHI Isa et OLM Christine, 2004, « L'observatoire des entrées et sorties du RMI à Paris » *in* : ARDILLY Pascal (dir.), *Echantillonnage et méthodes d'enquêtes*, Actes du 3ème colloque francophone sur les sondages de la Société Française de Statistique, Dunod.

BEYEME-ONDOUA Jean-Paul, 2007, « Evaluation de la qualité des données chaînées du cancer colorectal », *Santé publique*, vol 19, n° 6.

CNIL, 2001, *21ème rapport d'activité 2000. Délibération n°00-031 du 25 mai 2000*, La Documentation française.

FOURNEL Isabelle, « Détermination du statut vital par chaînage entre des données hospitalières et les données de mortalité nationales anonymisées », à paraître dans la *Revue d'épidémiologie et de santé publique*, n° 55.

GENSBITTEL Michel-Henry, RIANDEY Benoît, QUANTIN Catherine, 2007, « Appariements sécurisés : statisticiens, ayez de l'audace ! », *Courrier des statistiques,* n°121-122 (www.insee.fr).

GOUYON Béatrice, METRAL Pierre et al., 1999, « Réseau périnatal de Bourgogne », *Technologie de la santé*, 37, pp 51-56.

JARO Matthew A., 1995, « Probabilistic-linkage of large public health data files », *Statistics in Medecine,* 14, pp.491-498.

MIZRAHI Andrée et Arié, 2006, « Premiers sondages français dans les dossiers de sécurité sociale et appariement avec les enquêtes auprès des ménages », *in* : LAVALLEE Pierre et RIVEST Louis-Paul (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, Dunod.

QUANTIN Catherine, BOUZELAT Hocine, ALLAËRT François-André et all., 1998, « Automatic record hash coding and linkage for epidemiological follow-up data confidentiality », *Methods of Information in medecine*, 33, pp 271-277.

QUANTIN Catherine, GOUYON Béatrice, ALLAERT François-André, COHEN Olivier, 2006, « Proposition d'un identifiant à composante familiale rendu anonyme », *in* : LAVALLEE Pierre et RIVEST Louis-Paul (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, Dunod.

TARDIF Laurent, 2007, « Etude méthodologique du chaînage des séjours pmsi mco 2004 », *Documents de travail*, N° 116 – octobre, DREES, Ministère de la santé.

THIRION Xavier, SAMBUC Roland, SAN MARCO Jean-Louis, 1988, « Epidemiology and anonymity : a new method », *Revue d'épidémiologie et de santé publique*, 36, pp. 36-42.

TROUESSIN Gilles, ALLAERT François-André, 1997, « FOIN : a nominative information occultation function », *MIE*, 3, pp 196-200.

Liste des sigles

AGDREF : Application de gestion des dossiers des ressortissants étrangers en France

ANAEM : Agence nationale d'accueil des étrangers et des migrations

CepiDc : Centre épidémiologique sur les causes médicales de décès

CNIL : Commission nationale de l'informatique et des libertés

CHU : Centre hospitalier universitaire

CREDOC : Centre de REcherche et de DOcumentation sur les Conditions de vie

FAERE : Fichiers Anonymisés d'Elèves pour la Recherche et les Etudes

INE : Identifiant national des élèves/étudiants

INED : Institut national d'études démographiques

INSEE : Institut nationale de la statistique et des études économiques

INSERM : Institut national de la santé et des recherches médicales

NIR : Numéro d'inscription au répertoire (d'identification des personnes physiques : RNIPP), connu sous les appellations « n° de sécu », « n° INSEE »

ONED : Observatoire national de l'enfance en danger

RMI : Revenu minimum d'insertion

RNIAM : Répertoire national inter-régime de l'assurance maladie

SHA : du nom de Adi Shamir, l'un des promoteurs de la cryptographie, inventeur avec Ron Rivest et Len Adleman du système de cryptage à clé publique RSA (d'après leurs initiales).

SISE : Système d'information sur les étudiants

VIH : Virus d'immuno-déficience humaine

BENOÎT RIANDEY

Méthode sécurisée de traitement des fichiers administratifs

Les démographes locaux souhaitent accéder à des informations individuelles issues d'institutions distinctes et à pouvoir les apparier, mais se heurtent à l'obstacle tout à fait légitime de la confidentialité. Pas de transfert de données sans anonymisation. Pas d'appariement sans données nominatives. Justement, si ! Les épidémiologistes ont eu à mettre au point des méthodes d'appariements sécurisés par hachage des identifiants qui résolvent cette contradiction. Il appartient aux démographes de s'en emparer. Cet article présente brièvement le principe de deux types de ces méthodes, l'appariement déterministe fondé sur un identifiant simple et l'appariement probabiliste qui tend à combler l'absence d'identifiant ou sa mauvaise qualité. Quelques exemples effectifs ou potentiels d'applications à la démographie locale sont présentés.