L'accès aux données confidentielles de la statistique publique

De la sensibilité des données économiques à la sensibilité des données de santé

Jean-Pierre LE GLÉAU
Inspecteur général honoraire de l'Insee

En France comme dans l'ensemble de l'Europe, les données de la statistique publique font l'objet de dispositions légales particulières, qui viennent s'ajouter aux protections prévues par le droit pour toutes les catégories de données. Le « secret statistique » apparaît ainsi comme une garantie donnée aux répondants en contrepartie de la sincérité des informations qu'ils fournissent à l'administration. La loi prévoit des exceptions à ce secret pour la recherche scientifique : afin de répondre à la demande croissante d'accès à des données confidentielles de la part des chercheurs, des dispositifs spécifiques ont été mis en place ces dernières années. Ces modalités déjà éprouvées peuvent, peut- être, servir de modèle pour faciliter l'accès des chercheurs à d'autres types de données sensibles, telles que les données de santé.

Pourquoi le secret statistique ?

Pour connaître l'économie et la démographie d'un pays, il est nécessaire d'interroger de temps à autre ses habitants, ses entreprises, ses exploitations agricoles, ses administrations. Toutes les informations ainsi recueillies permettent, combinées à d'autres, de dresser un tableau aussi fidèle que possible de la réalité du pays.

Si l'on veut que l'information recueillie soit utile pour dresser un panorama conforme à la réalité, il faut que celui qui est interrogé (particulier ou entreprise) fournisse une réponse sincère. Pour cela, il est indispensable qu'il soit assuré que ses réponses ne seront pas utilisées dans un sens qui puisse lui causer du tort. On pense immédiatement aux impôts, à la police, mais aussi aux concurrents, aux voisins, à la famille, etc. Les réponses aux questions doivent donc rester absolument confidentielles afin de pouvoir donner au répondant l'assurance que sa réponse ne lui portera pas tort.

C'est justement l'objet du secret statistique : protéger les informations recueillies au moyen d'enquêtes statistiques, afin d'obtenir de la part de la personne interrogée des réponses sincères, tout en lui garantissant que ces réponses ne pourront en aucun cas lui porter préjudice.

Comme toutes les informations individuelles détenues par l'administration, celles qui sont recueillies au moyen d'enquêtes statistiques sont protégées par des règles générales de confidentialité qui s'appliquent pour tous les agents ayant à en connaître. Cependant, le secret statistique va plus loin, car il confine les données ainsi recueillies dans une "bulle" très étanche, interdisant par exemple leur communication d'une administration à une autre.

Les dispositions légales

Le secret statistique est présent dans le traité de l'Union Européenne, qui précise dans son article 338 que « L'établissement des statistiques se fait dans le respect (…) de la confidentialité des informations statistiques ».
Il est développé dans le règlement de mars 2009 relatif aux statistiques européennes, qui lui consacre un chapitre (articles 20 à 26) dans lequel sont énumérées les mesures s'appliquant « pour garantir que les données confidentielles sont utilisées exclusivement à des fins statistiques et pour empêcher leur divulgation illicite »
Par ailleurs, la loi statistique française du 7 juin 1951, maintes fois modifiée, définit dans ses articles 6 et 7bis la protection qui doit être apportée aux informations couvertes par le secret statistique.
Mais quelles sont au juste ces informations couvertes par le secret statistique ?
Les informations couvertes par le règlement de l'Union européenne sont celles qui sont utilisées par cette dernière pour produire les résultats demandés par les instances européennes.
Celles qui sont protégées par la loi française sont de deux types :

- d'une part, les informations collectées au moyen d'enquêtes statistiques, ayant reçu un numéro de visa attribué par le ministre chargé de l'économie ; leur liste est publiée chaque année au Journal officiel ;
- d'autre part, les informations recueillies par l'Insee ou des services statistiques ministériels (SSM) auprès d'autres administrations, en vue d'établissement de statistiques.

Pour la France, il y a un très fort recouvrement entre ces deux ensembles.
Le règlement européen et la loi française fixent un principe général, des modalités de mise en œuvre, ainsi que quelques exceptions :

- le principe est celui d'un confinement des informations confidentielles auprès de celui qui les a collectées et traitées pour produire des statistiques anonymes : il ne doit les communiquer à personne, pas même à d'autres administrations ;
- en France, ces informations restent secrètes pendant un temps assez long (25 ans pour les informations d'ordre économique et financier, 75 ans pour celles qui ont trait aux faits et comportements d'ordre privé) et de lourdes sanctions s'appliquent à qui trahirait ce secret (un an de prison et 15 000 € d'amende) ;
- des exceptions sont prévues dans la loi française. Certaines sont d'ordre institutionnel, d'autres visent à favoriser l'emploi des données collectées à des fins de recherche scientifique. Les exceptions institutionnelles visent par exemple à obliger quiconque aurait connaissance d'un crime ou d'un délit grâce à ces données confidentielles, à en informer le procureur ; elles permettent aussi à ce dernier d'avoir connaissance d'informations couvertes par le secret statistique, dans le cadre d'une commission rogatoire. Ces dispositions sont une entorse au secret statistique, peu conformes aux principes européens. Par ailleurs, une disposition permet aux chercheurs d'avoir accès à la masse des informations collectées en vue de l'établissement des statistiques. Ce dispositif est prévu aussi bien dans le cadre européen que dans celui de la loi française.

L'accès des chercheurs aux informations confidentielles

Au moment où la loi statistique française a été adoptée (1951), la question de l'accès des chercheurs aux données individuelles ne se posait guère. Tous les questionnaires étaient sur support papier, et il était hors de question de permettre un accès général, fût-ce à des fins de recherche, à ces documents. Seules des dérogations ponctuelles, très rares, étaient envisageables.
Avec le développement de l'informatique, tant du côté de la recherche que du côté des services producteurs de statistiques, le contexte a changé.

Il est devenu possible de mettre à la disposition des chercheurs des fichiers de données individuelles rendues anonymes. En supprimant ou agrégeant un certain nombre de variables, on pouvait constituer des fichiers présentant un intérêt pour la recherche, mais préservant l'anonymat de ceux qui avaient répondu.

C'est ainsi qu'ont été créés et diffusés des fichiers détail permettant aux chercheurs de travailler sur des données individuelles, sans porter atteinte à la vie privée des personnes qui avaient répondu aux enquêtes.

Malheureusement, ces fichiers individuels ne pouvaient être produits pour les entreprises. En effet, dès lors que l'on donne pour une entreprise son activité économique, sa taille, voire un indice sur sa localisation, il devient souvent très facile de deviner de quelle entreprise il s'agit. On a donc été obligé de constater que la plupart des fichiers individuels d'entreprises ne permettaient pas de sauvegarder le secret statistique. Quel serait, pour un chercheur, l'intérêt d'un fichier individuel d'entreprises, dans lequel on aurait fait disparaître les variables « activité économique », « taille » et « localisation » ?

En 1984, la loi sur le secret a donc été modifiée, afin de permettre, sous condition, l'accès des chercheurs aux informations individuelles sur les entreprises. Cela revient à élargir aux chercheurs la « bulle » dans laquelle étaient jusqu'alors confinées ces informations.
La loi prévoit que cet élargissement peut se faire après avis d'un comité créé à cet effet et appelé « Comité du secret statistique concernant les entreprises ». Il était présidé par un membre du Conseil d'État et comprenait quatre représentants de l'administration (dont un représentant du ministre de la Justice), quatre représentants des entreprises, un représentant des organisations syndicales de salariés et un représentant des utilisateurs régionaux et locaux de la statistique publique.

Après avis de ce comité, les chercheurs recevaient les informations demandées sur support magnétique. Ils avaient auparavant signé un engagement de confidentialité et s'étaient engagés à détruire les données à un certain terme fixé par le comité du secret statistique.

Pour les fichiers de données sur les ménages, la mise à disposition de fichiers individuels a donné satisfaction pendant quelque temps. Mais il est vite apparu deux inconvénients :

- d'une part, avec le développement d'internet, ces fichiers étaient accessibles sur le site de l'Insee et donc n'importe qui pouvait les télécharger, de façon anonyme. Y compris des personnes, éventuellement animées d'intentions malveillantes, et connaissant par exemple certains répondants à l'enquête. Au moyen des informations déjà connues sur ceux-ci, il leur était dans certains cas possible d'identifier l'enregistrement correspondant et donc de prendre connaissance des réponses effectuées par ces personnes. Ces cas étaient très rares, supposaient de la part de l'internaute une démarche volontaire et complexe ainsi qu'une véritable envie de violer la loi, mais ils représentaient un danger trop grand pour la préservation du secret statistique ;
- d'autre part, pour assurer une protection élémentaire du secret, ces fichiers ne comportaient pas tout le détail qui aurait été nécessaire aux chercheurs : profession codée sur deux chiffres et non sur quatre, localisation à la région seulement, etc.

Pour répondre à ces inconvénients, une première démarche a consisté à mettre à disposition des chercheurs, et d'eux seulement, des fichiers un peu plus détaillés, mais préservant encore l'anonymat, pour qui ne tenterait pas systématiquement d'identifier des individus. C'est ce que l'on appelle les « Fichiers de production et de recherche » (FPR), mis à la disposition des chercheurs, via le réseau Quetelet. Le réseau Quetelet s'assure que le demandeur est bien un chercheur et lui donne accès à des fichiers « raisonnablement anonymes », c'est-à-dire où il n'est pas possible d'identifier qui que ce soit, tant que l'on utilise ces fichiers à des fins de

recherche scientifique. Un chercheur qui romprait cet engagement pourrait tenter d'identifier une ou deux personnes dont il saurait qu'elle a participé à l'enquête. Le plus souvent il n'y arriverait pas. Mais, exceptionnellement, il pourrait y parvenir. Le chercheur serait alors en infraction avec la loi et, compte tenu de la traçabilité du réseau Quetelet, il courrait un risque sérieux d'être démasqué.

Mais, avec le développement des nouveaux moyens (logiciels et matériels) de traitement de données, les chercheurs ont eu besoin d'une information encore plus détaillée pour leurs travaux.

C'est pourquoi la loi a été à nouveau modifiée en 2008 pour permettre l'accès aux données les plus détaillées sur les ménages, à des fins de recherche scientifique ou historique, ou de statistique publique. Le comité du secret statistique doit donner son avis pour une telle communication. Dans ce but, il a été renommé et réorganisé, pour permettre l'entrée d'organisations concernées par la transmission de données à caractère personnel. En effet, ces données permettant parfois l'identification des personnes, il est nécessaire, pour y avoir accès, d'accomplir également des formalités auprès de la Cnil, variables selon la nature des informations concernées.

Comment accéder aux données ?

La procédure formelle d'autorisation inclut l'avis du comité du secret statistique, l'accord de l'autorité dont émanent les documents (le plus souvent l'Insee ou un service statistique ministériel) puis une décision de l'administration des archives, puisque les enquêtes statistiques sont considérées comme des archives publiques.

Munis de ces autorisations, les chercheurs peuvent donc avoir accès à l'ensemble des données recueillies ou traitées par la statistique publique. C'est un progrès majeur. Mais cette facilité impose en contrepartie une protection supplémentaire de ces données.

En particulier, il était apparu depuis quelque temps que la procédure qui permettait à des chercheurs d'emporter dans leur labo un CD comportant des données confidentielles ne présentait pas les garanties suffisantes en termes de protection du secret statistique.

C'est pourquoi, l'Insee a développé, grâce au Groupement des écoles nationales d'économie et de statistique (Genes), un centre d'accès sécurisé aux données, le CASD.

Placé sous la responsabilité du Genes, le CASD permet aux chercheurs qui ont été habilités par le comité du secret statistique d'accéder aux données les plus confidentielles. Ils ont signé un engagement de confidentialité et la loi leur interdit toute communication externe, sous peine des sanctions mentionnées précédemment.

Le CASD leur remet alors une boîte appelée « SD Box », qui est un terminal grâce auquel les chercheurs peuvent accéder au serveur, localisé au Genes, sur lequel se trouvent les données confidentielles. Le système leur permet de « voir » les données, de travailler dessus, mais en aucun cas de les imprimer, ou de les recopier sur un autre support (clé USB, disque dur, etc.). Lorsque les chercheurs ont obtenus les résultats qu'ils souhaitaient, ils les placent dans une boîte à lettres virtuelle. Des experts du CASD vérifient que ces résultats ne contreviennent pas aux règles du secret statistique. Si tel est le cas, le fichier de résultat est renvoyé au chercheur par simple messagerie.

Ce système permet une bonne protection du secret statistique, en assurant la traçabilité de toutes les opérations réalisées par les chercheurs. Mis en place depuis 2010, il semble aussi donner satisfaction aux chercheurs qui peuvent enfin avoir accès à l'information la plus détaillée.

Un exemple pour l'accès à d'autres types de données confidentielles ?

Les données statistiques sont cependant loin d'être les seules données publiques couvertes par une obligation de confidentialité.

On peut en citer bien d'autres. Pour certaines d'entre elles, l'exemple de l'accès aux données de la statistique publique peut être riche d'enseignements.

Ainsi, la loi du 22 juillet 2013 permet, sous certaines conditions, l'accès des chercheurs aux données fiscales. Celui-ci devrait se faire, après avis favorable du comité du secret statistique, par l'intermédiaire d'un centre d'accès sécurisé, qui permet de préserver la confidentialité et d'empêcher la dissémination des informations par inadvertance ou malveillance. Le centre d'accès sécurisé aux données de la statistique publique semble avoir toutes les qualités requises pour servir d'instrument de mise à disposition des données fiscales. On aurait ainsi, pour les données fiscales, un cheminement analogue à celui qui existe pour les données de la statistique publique : avis du comité du secret statistique, puis accès par un centre d'accès sécurisé. La seule différence serait que l'accord de l'autorité dont émanent les documents et la décision des archives serait remplacés par une décision du ministre chargé du budget.

On peut imaginer un processus analogue pour l'accès à des données d'ordre médical. Les traitements de données à caractère personnel ayant pour fin la recherche dans le domaine de la santé font l'objet d'un chapitre particulier (Chapitre IX, articles 53 à 61) de la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Le législateur a considéré que les données relatives à la santé devaient bénéficier d'une protection particulière : elles sont en effet classées parmi les données sensibles dans cette même loi, au même titre que celles qui font apparaître les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou qui sont relatives à la vie sexuelle de celles-ci.

Il convient donc de distinguer de façon précise les données médicales selon qu'elles permettent ou non d'identifier, directement ou indirectement, les individus auxquels elles se rapportent. Ce travail aboutit généralement à la classification des données médicales en trois catégories :

- celles qui sont totalement anonymes et ne permettent aucune identification ;
- celles qui permettent, de façon directe ou, le plus souvent, indirecte l'identification de certains individus ;
- celles qui se trouvent dans une zone intermédiaire : elles ne permettent en général pas une identification des personnes concernées, mais, certains individus possédant des informations spécifiques pourraient, en se donnant du mal, avoir une probabilité non négligeable d'identifier un petit nombre d'individus.

Ces trois catégories doivent être traitées de façon spécifique.

La première contient un grand nombre de tableaux statistiques, où on s'est assuré qu'il y avait un nombre suffisant d'individus dans chaque case. Elle contient aussi des fichiers de données individuelles, où les variables liées à chaque personne ont été suffisamment agrégées ou floutées pour rendre impossible toute identification. Ces tableaux et ces fichiers peuvent être mis sans inconvénient à disposition du public le plus large (internet, open data,...)

La troisième catégorie constitue ce que l'on appelle parfois la « zone grise ». L'identification d'individus de la base n'est pas strictement impossible, mais elle nécessiterait la disposition d'informations complémentaires faiblement répandues dans le public, le déploiement de moyens importants pour tenter de parvenir à une identification, avec comme résultat une simple probabilité, sans certitude, d'avoir identifié quelques individus (en général peu nombreux). La mise à disposition de tels fichiers doit être rendue possible, pour des personnes dont on aurait vérifié le sérieux et la moralité, après qu'elles auraient détaillé le projet pour lequel elles ont besoin d'avoir accès à ces données. Ces personnes signeraient un engagement de n'utiliser ce fichier que dans le cadre dudit projet et de ne tenter en aucun cas d'identifier un individu précis de la base. Tout manquement à cet engagement a des chances non négligeables d'être repéré et de lourdes sanctions peuvent s'appliquer dans ce cas. Cette catégorie s'apparente à celle qui a été décrite pour l'accès aux données de la statistique publique, via le réseau Quetelet.

La deuxième catégorie nécessite un contrôle beaucoup plus serré, car elle est constituée de fichiers permettant, à un faible coût et sans disposer d'une information rare, d'identifier un grand nombre d'individus. L'accès à ces données pourrait se faire par l'intermédiaire d'un centre d'accès sécurisé, avec toutes les garanties déjà prévues pour les données issues de la statistique publique et, bientôt, pour les données fiscales :

- présentation d'un projet pour lequel il est démontré qu'il est nécessaire d'avoir accès à des informations très détaillées et donc potentiellement identifiantes ;
- garanties sur la personne qui présente la demande : environnement institutionnel, caution d'une personne d'un rang hiérarchique suffisant, accès dans un environnement matériel correctement sécurisé ;
- avis d'une autorité (analogue au comité du secret statistique) sur le respect de ces critères et formalités adaptées auprès de la Cnil ;
- identification forte (biométrique) de la personne ayant accès aux données ;
- consultation et travail rendus possibles sur des données ; mais celles-ci restent sur un serveur et ne peuvent être ni copiées, ni transcrites, ni imprimées sur aucun support (papier, CD, clef USB…) ;
- vérification par des experts que les résultats issus de ces travaux sont strictement anonymes.

Un organisme rendant ce genre de service pourrait être du même type que le centre d'accès sécurisé aux données de la statistique publique (CASD).

Des sanctions pénales restent prévues dans ce cas, mais l'idée est plutôt de rendre quasiment impossible l'accès aux données personnelles et de repérer facilement ceux qui auraient tenté d'enfreindre les interdictions fixées par la loi.

Toutes ces conditions permettent des modalités d'accès adaptées au degré de confidentialité des données, qui préservent la vie personnelle des individus tout en permettant un travail de recherche dans de bonnes conditions.