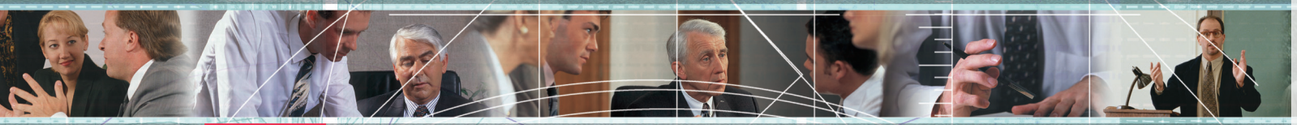

les dispositifs d'évaluation des compétences et les comptes de l'éducation. Dans ce cadre, elle a mis en place notamment un système d'évaluation et d'étalonnage des établissements scolaires, visant à remplacer les « palmarès »peu rigoureux diffusés par certains magazines.

La notion de **système d'informations partagées** est importante dans le domaine de l'éducation. D'une part les établissements ont des degrés de tutelles différents vis-à-vis du ministère. La tutelle est totale pour les établissements du premier et du second degré. En revanche, dans l'enseignement supérieur, il convient de « convaincre ».D'autre part, du fait des lois successives de décentralisation, les collectivités locales ont d'importantes responsabilités sur les écoles, les collèges et les lycées. La DEPP développe et négocie des nomenclatures partagées. Elle produit des données décrivant le système éducatif, la réalité des parcours et l'efficacité de certaines mesures. De plus, l'introduction de la culture de la performance par la Lolf a en partie changé la donne. Pour qu'une évaluation de la performance du système dans la durée soit possible, il importe que les investigations ne soient pas entièrement tributaires d'un système d'information susceptible d'être modifié pour des raisons administratives. Il convient donc d'articuler et de compléter la production standard de l'information avec des enquêtes adéquates.

Faciliter l'accès aux sources administratives

La loi de 1951 relative à l'obligation, à la coordination et au secret en matière de statistiques a été révisée en 2004. En particulier l'article 7 *bis*, dans cette nouvelle rédaction, stipule que « Sur demande du ministre chargé de l'économie, après avis du Conseil national de l'information statistique, et sauf disposition législative contraire, les informations relatives aux personnes physiques, à l'exclusion des données relatives à la vie sexuelle, et celles relatives aux personnes morales, recueillies dans le cadre de sa mission, par une administration, une personne morale de droit public, ou une personne morale de droit privé gérant un service public sont cédées, à des fins exclusives d'établissement de statistiques, à l'Institut national de la statistique et des études économiques ou aux services statistiques ministériels ».
Ainsi, si auparavant, les administrations **pouvaient** transmettre des données individuelles aux services statistiques publics (sous réserve du respect des règles de protection de confidentialité), elles y sont désormais **tenues** si ces derniers les leur demandent. Cette modification vise à faciliter l'utilisation des sources administratives.

Très anciennes, les statistiques de la **justice** dépendent, elles aussi, du fonctionnement de l'institution. Elles n'ont parfois qu'un rapport indirect avec les phénomènes de société, tels que la délinquance, qui conduisent à mettre en œuvre les justices pénale et civile. Pour cette raison, il a été jugé utile d'organiser des « enquêtesde victimation », afin d'appréhender l'insécurité ressentie par les personnes. Par ailleurs, le fait que les statistiques soient des sous-produits de l'activité des tribunaux et qu'elles soient utilisées comme indicateur de performance de ceux-ci, peut avoir des effets indirects fâcheux. En effet, certains acteurs peuvent être incités à infléchir leurs activités en fonction du contenu et des modalités de la quantification induite par ces indicateurs. Ce risque existe dans toute organisation, dès lors que les indicateurs n'ont pas seulement un rôle d'information, mais contribuent à l'évaluation des acteurs et influencent donc leur comportement.

Le recours aux enquêtes demeure indispensable

Les sources administratives et plus généralement les systèmes d'information des grands organismes publics et privés, constituent une ressource irremplaçable pour la statistique future. Cependant, le propre de ces systèmes est de refléter les institutions et leurs actions, telles qu'elles sont mises en formes et codifiées à un moment et dans un contexte juridique et culturel spécifique. Les phénomènes émergents de la société ne peuvent pas être décrits et quantifiés par ces systèmes alors qu'ils peuvent être saisis par des enquêtes. Elles offrent en effet une grande latitude dans l'élaboration des questionnaires, alors que les fichiers administratifs ne font que suivre, avec un certain retard, l'évolution de ces phénomènes. Le recours à des enquêtes directes reste donc indispensable pour appréhender les questions sociales. Les enquêtes de victimation en sont un bon exemple. Des questions comme celles de la pauvreté ou des sans-abri ne peuvent être traitées uniquement par des sources émanant des fichiers des organismes d'aide à ces personnes. Seul le rapprochement des deux types d'informations permet d'éclairer ces situations.

Plus généralement se pose la question de la **mise en catégorie**. Les équivalences conventionnelles nécessaires au comptage sont, dans le cas des sources administratives, antérieures au travail du statisticien et induites par des logiques diverses, liées au droit ou aux besoins de l'action. Il est indispensable que le chercheur, et notamment le statisticien, puisse mettre en œuvre d'autres hypothèses, d'autres découpages ou équivalences possibles. Tous les progrès récents de la statistique publique ont pris appui sur la dualité de ces approches, l'une permettant de décrire les institutions telles qu'elles existent et l'autre d'explorer, au moins en partie, la société telle qu'elle n'est pas encore complètement instituée, afin de rendre éventuellement possibles et pensables de nouvelles façons de faire.

Pour en savoir plus :

- Desrosières A. : « Enquêtes versus registres administratifs : réflexion sur la dualité des sources statistiques », *Courrier des statistiques* n° 111, septembre 2004.

CNIS
CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE
Secrétariat Général du CNIS
Timbre D130,
18 boulevard Adolphe Pinard,
75 675 Paris Cedex 14
Téléphone : 01 41 17 52 62
Télécopie : 01 41 17 55 41
www.cnis.fr

N° 5
Décembre 2006
Directeur de la publication :
Jean-Pierre PUIG
Rédacteur en chef :
Carla SAGLIETTI
Responsables éditoriales :
Brigitte OUVRÉ, Anne DOLEZ
Maquette : STE
Publication diffusée gratuitement,
ne peut être vendue

CNIS
CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE
Chroniques

Enquêtes statistiques et sources administratives : une complémentarité à exploiter

Dès leur origine, les statistiques publiques se sont abreuvées à deux types de sources. Les enquêtes, exhaustives telles que le recensement de population, ou par sondage comme l'enquête sur l'emploi, sont complémentaires des sources administratives telles que les registres de l'état civil, les fichiers de l'ANPE, etc. Les premières fournissent directement les informations jugées utiles et pertinentes, grâce aux techniques spécifiques du métier de statisticien, notamment celle de l'échantillonnage. Les secondes présentent l'avantage d'être déjà disponibles. Elles sont donc *a priori* susceptibles d'être moins coûteuses et elles allègent la charge pesant sur les répondants. De plus, à la différence des enquêtes par sondage, elles sont souvent quasi exhaustives, ce qui permet de les mobiliser pour fournir des informations régionales et locales. Mais elles présentent un inconvénient : les données qu'elles contiennent sont recueillies dans un objectif de gestion, c'est-à-dire avec des finalités et selon des procédures différentes de celles que nécessite une information statistique pertinente. Tels sont les principaux termes d'un débat récurrent, qui parcourt toute l'histoire de la statistique publique.

Depuis le début des années 1990, les éléments de ce débat ont évolué, pour des raisons tenant à la fois aux transformations des **usages sociaux** de la statistique et **aux progrès constants et rapides de l'informatique**. D'une part, la conception, la mise en place et surtout l'évaluation des politiques publiques impliquent un recours accru à des méthodes reposant sur l'analyse statistique de fichiers de **données individuelles** et plus seulement sur des « données globales ». D'autre part, les organismes privés ou publics raisonnent désormais davantage en termes de **systèmes d'information** que de registres administratifs. Ces différentes évolutions, tant sociales que techniques, ont fait l'objet d'un large débat, lors de l'Assemblée plénière du Cnis tenue le 29 novembre 2005, à partir des exemples des statistiques de la santé, de l'éducation et de la justice.

L'argumentaire classique qui met en balance les avantages et les inconvénients des deux types de sources a donc partiellement changé. L'intérêt principal des données issues des fichiers de gestion des administrations et des entreprises est qu'elles sont désormais coordonnées et structurées dans des systèmes d'information qui incluent d'emblée leurs diverses finalités. En effet, ces derniers sont le fruit d'une collaboration entre les comptables, les responsables des relations humaines, tous les fournisseurs et les utilisateurs des informations, ainsi que les statisticiens et les informaticiens. Ils produisent d'abondantes données individuelles ou déjà agrégées, qui sont déposées et mises en forme dans des **entrepôts de données,** incomparablement plus riches que ne l'étaient les anciennes sources administratives. C'est désormais parfois le « trop plein » qui pose problème : comment explorer ces gisements d'informations et en extraire les éléments et les synthèses pertinentes pour les utilisateurs ? Des techniques nouvelles dites de *data mining* proposent des outils de statistique exploratoire pour naviguer dans ces univers.

Un autre avantage de ces nouveaux systèmes est qu'ils évitent la multiplication des interrogations redondantes des personnes et des entreprises. La charge des enquêtes et les refus de réponse sont en effet des préoccupations partagées par tout le système statistique. Les deux types de sources apparaissent ainsi plus complémentaires qu'antagonistes. Par exemple, les données recueillies dans les enquêtes auprès des entreprises proviennent souvent de leurs systèmes d'information. La distinction classique entre les deux types de sources n'est, dans ce cas, plus aussi pertinente qu'elle ne l'est pour les statistiques portant sur les individus et les ménages.

CNIS

CONSEIL NATIONAL DE L'INFORMATION STATISTIQUE

Les questions soulevées par l'usage statistique des fichiers administratifs

Malgré tous les apports du recours aux systèmes d'information, des interrogations demeurent quant à cette façon nouvelle de penser la statistique publique dans la société. Elles concernent l'encadrement juridique, la qualité des données, le retraitement statistique, la difficulté de réaliser des comparaisons et la question de savoir ce que l'on mesure exactement.

Un **cadre juridique** est nécessaire pour garantir la possibilité d'effectuer des appariements de fichiers. La mobilisation et la réutilisation de données individuelles collectées à des fins de gestion sont strictement encadrées par des règles visant à protéger la confidentialité et la vie privée des personnes. La Commission nationale de l'informatique et des libertés (Cnil) veille à l'application de ces règles. Les projets portant sur de nouveaux usages de données administratives doivent lui être soumis. Dans le cas des enquêtes directes, les personnes interrogées doivent être informées du traitement qui sera appliqué à leurs réponses, en particulier si celui-ci implique un appariement avec des fichiers administratifs. Les traitements portant sur les données dites « sensibles »: origines ethniques, opinions politiques, philosophiques ou religieuses, appartenance syndicale, santé, vie sexuelle, sont également réglementés. Ainsi, l'accès des statisticiens aux données relatives à la santé est désormais possible après autorisation de la Cnil.

Une autre interrogation porte sur la **qualité des données,** nécessaire à la technique professionnelle des statisticiens. Malgré les efforts des statisticiens et des informaticiens pour assurer la cohérence d'ensemble des systèmes, il est courant que les questions et les variables directement utiles à la gestion soient beaucoup mieux renseignées et contrôlées avec plus de soin et d'exhaustivité que les autres. Elles sont donc de meilleure qualité que celles dont la présence dans le fichier n'est due qu'à la demande des statisticiens. Ainsi se pose la question de la participation des statisticiens, au sein des administrations concernées, à l'élaboration des systèmes d'information mais aussi au fonctionnement de ces systèmes au-delà de leur conception.

Le retraitement des fichiers administratifs à des fins statistiques est coûteux, en argent, en temps de travail et en matière grise. Cette utilisation nécessite en fait un lourd investissement pour le statisticien qui doit s'approprier le fichier par l'expertise de la qualité des informations. Les statisticiens ne se contentent pas de reproduire des tabulations effectuées par d'autres quand ils travaillent sur des fichiers administratifs. Ils traitent, interprètent et transforment à leur façon le matériau fourni, ce qui, là encore, rend indispensable une grande proximité du statisticien avec les producteurs du fichier.

Le recours à des sources administratives peut rendre difficiles les **comparaisons** intertemporelles et internationales. En effet, la coordination d'un système d'information dans un cadre institutionnel précis, par exemple national, est un avantage. Mais cet avantage peut se retourner et devenir un inconvénient. C'est le cas pour les séries temporelles quand ce cadre est modifié. En particulier les changements de réglementation perturbent la continuité des séries comme, par exemple, le suivi des séries de demandeurs d'emploi lors du transfert des inscriptions de l'ANPE vers l'Assedic. Le problème se pose également pour élaborer des comparaisons internationales. Celles-ci sont de plus en plus prisées, notamment dans le cadre des politiques européennes incitatives reposant sur l'étalonnage (ou *benchmarking*) et la comparaison des performances nationales. Certains classements reposent sur des données issues de sources différentes : registres administratifs pour certains pays (notamment nordiques), enquêtes directes pour d'autres. C'est ainsi que sont élaborées, par exemple, les statistiques de l'emploi et du chômage, ou l'enquête européenne sur les revenus et la pauvreté (enquête SILC). D'ailleurs l'harmonisation européenne pousse à développer des enquêtes, puisqu'il est de plus en plus difficile de fournir des informations comparables à partir de sources administratives nationales hétérogènes.

Enfin, la confrontation des deux types de sources pose la question épistémologique, sinon philosophique, de **ce qui est au juste quantifié par le recours aux fichiers administratifs**. S'agit-il directement de phénomènes de société : santé, niveau d'éducation, délinquance… ? Ou plutôt des moyens dont la société se dote pour les identifier, les qualifier et agir sur eux : le système hospitalier, les écoles, la police et les tribunaux ? Cette question est fréquemment soulevée, par exemple à propos de la pauvreté, du chômage ou de la criminalité, dont la « mesure »est souvent issue de sources administratives. Telle « évolution » peut ainsi être interprétée, soit comme celle du phénomène lui-même, soit comme celle de la mesure de l'activité du service chargé d'évaluer et de gérer ce phénomène. Ceci est particulièrement vrai quand une problématique commence à être perçue comme un « problème social ». Dans ce cas, les moyens d'action et d'observation sont mis en place en même temps et la statistique reflète cette double évolution simultanée.

Cette ambiguïté est source de malentendus dans le débat public. Les controverses sur la « nouvellepauvreté »dans le cadre de la mise en place du RMI ou de la CMU, ou celles sur les violences subies par des femmes et des enfants lors de la création de « numéros verts » en sont des exemples.

Les statistiques d'entreprises à l'origine de l'utilisation des sources administratives

L'Insee utilise des sources administratives pour faire des statistiques depuis plusieurs décennies. Historiquement, c'est le domaine des statistiques d'entreprises qui a été le terrain des premières expériences. Ces expériences ont conduit à développer une stratégie visant à ne procéder à des enquêtes statistiques que lorsque les sources administratives ne permettent pas d'obtenir les résultats statistiques souhaités. Un des éléments de cette stratégie a consisté en la création et la gestion par l'Insee, il y a plus de 30 ans, d'un répertoire de l'ensemble des entreprises françaises (Sirene), intégrant la notion d'identifiant unique qui permet, en théorie, des appariements aisés entre les sources qui l'utilisent.

Le contexte nouveau de l'évaluation et de l'étalonnage

Les demandes adressées à la statistique publique ont changé. Longtemps, elles ont visé à dresser des tableaux décrivant l'état de la nation au niveau macroéconomique. Elles sont maintenant de plus en plus destinées à poser des diagnostics précis sur des problèmes sociaux pointus ou, compte tenu des lois de décentralisation successives, à des niveaux géographiques fins : région, département, commune... Ceux-ci correspondent à des politiques spécifiques définies territorialement, comme par exemple les zones urbaines sensibles (Zus), ou à des politiques régionales européennes. Ces diagnostics visent à la fois à cerner un problème, à identifier les éventuels leviers d'action et enfin à proposer des outils d'évaluation et d'étalonnage de ces politiques, outils s'appuyant souvent sur des méthodes économétriques. Les informations étayant ces diagnostics et ces évaluations sont considérées par leurs usagers comme des **indicateurs**. Ce terme, de plus en employé, résume bien ce style d'action publique.

Ce vocabulaire et cette façon d'utiliser les statistiques se retrouvent par exemple dans la **Loi organique relative aux lois de finances** (Lolf), votée en France en 2001, ainsi que dans les nouvelles formes d'organisation des politiques européennes, comme la **méthode ouverte de coordination** (Moc) explicitée au Sommet de Lisbonne en 2000. Le propre de ces méthodes est d'être très gourmandes en statistiques issues de sources différentes, enquêtes et fichiers. En effet, les études économétriques nécessitent des fichiers de données individuelles, souvent sous forme de panels longitudinaux. De même les résultats des politiques ciblées territorialement sont souvent présentés dans des tableaux d'indicateurs multiples. C'est typiquement le cas des politiques portant sur les quartiers situés dans des Zus, ou des politiques environnementales (indicateurs de développement durable).

Santé, éducation, justice : trois exemples de complémentarité des sources

Le projet actuellement porté par la direction de la Recherche, des Études, de l'Évaluation et des Statistiques (Drees) illustre, dans le domaine de la **santé publique**, ce que pourrait être une utilisation combinée de sources complémentaires, enquêtes et fichiers. La loi d'août 2004 sur la santé publique et l'assurance maladie a défini une centaine d'objectifs assortis chacun de quatre indicateurs. De plus, cette loi a initié un ensemble de réformes dont le suivi implique un grand nombre d'indicateurs supplémentaires. Des données individuelles sont ici nécessaires : les dépenses de soins sont souvent l'unique source du suivi de certaines pathologies. Ainsi, dans le cadre de ce projet, pourraient être mobilisés :

- les comptes nationaux, pour évaluer l'effort financier de la nation ;
- des enquêtes pour connaître les comportements des agents par catégories (CSP, revenus, état de santé) ;
- et enfin des bases de données administratives, comme le programme de médicalisation du système d'information (PMSI) qui fournit des données sur chaque séjour en hôpital et le système national d'informations inter régimes d'assurance-maladie (SNIIRAM) qui décrit l'ensemble des remboursements opérés dans le pays.

Ces deux bases, très riches, ne comportent cependant aucune information sur les caractéristiques des patients, ni sur leur état de santé, ce qui rend indispensable qu'elles soient complétées par des enquêtes spécifiques. Ce système d'information pourrait encore être enrichi par des données issues des systèmes d'assurances complémentaires, mutuelles et assurances privées : des réflexions sont en cours dans ce but. Les dépenses des ménages en soins de santé sont fortement déterminées par le niveau du « reste à charge », mais cette influence est encore mal connue et évaluée. Il y a là un cas typique où une bonne coordination, respectant les points de vue de chaque partie, est nécessaire. Les systèmes d'information doivent pouvoir être partagés, entre des acteurs autonomes et responsables, et non pas simplement décidés par une autorité centrale.

La statistique publique de l'**éducation** a connu des évolutions comparables. Sa direction a récemment complété son nom, devenant désormais la direction de l'Évaluation, de la Prospective et de la Performance (DEPP). De longue date, les enquêtes et les fichiers sur les élèves et les enseignants ont été étroitement associés, au point que, dans ce domaine, la dichotomie entre les deux types de sources peut sembler artificielle. Les données administratives sont adaptées en amont aux usages statistiques. La DEPP cherche à agir sur la définition des systèmes d'information, afin d'y inclure les variables d'intérêt statistique, par exemple la catégorie socioprofessionnelle des parents. Au sein du ministère, elle centralise les questions de nomenclatures et gère le répertoire des établissements. Elle participe à la gestion des systèmes statistiques académiques dont elle vérifie ainsi la qualité. Elle suit

L'utilisation des sources administratives dans les statistiques sociales

Les sources administratives présentent aujourd'hui une qualité très satisfaisante dans un certain nombre de domaines. La combinaison au niveau individuel de ces sources avec des données d'enquêtes se généralise dans le domaine des statistiques sociales.

Le meilleur exemple est sans doute celui des revenus. La source administrative fiscale offre aujourd'hui une qualité homogène et très satisfaisante en ce qui concerne les données sur les revenus. De plus, les analyses ont montré que les niveaux de revenus étaient systématiquement sous-évalués dans les enquêtes par rapport aux sources fiscales. En effet, les risques d'erreur dans les réponses à une enquête sont nombreux, comme les confusions entre les euros et les francs, mensuel/annuel ou simplement l'oubli de certaines sources de revenu. Par ailleurs, certains ménages se réfèrent à des documents, d'autres non.

Dans la source fiscale, le haut et le bas de la distribution des revenus sont toutefois susceptibles d'être de moins bonne qualité. Cette source est fondée sur le concept de revenu déclaré, parfois éloigné du concept économique de certaines composantes comme les revenus du patrimoine. La source administrative ne permet pas une mesure correcte de ces types de revenu. Il faut alors recourir aux données d'enquête.

Un élément qui plaide pour une utilisation accrue des sources administratives est l'impression désagréable qu'ont certains ménages de donner les mêmes informations à plusieurs administrations, d'autant que le nombre d'enquêtes a plutôt tendance à augmenter. La construction d'un système d'information doit donc utiliser au mieux les avantages comparatifs des sources administratives et des enquêtes et exploiter leur complémentarité, selon les objectifs visés.