Enjeux et opportunités de l'ouverture des données publiques en matière d'énergie

Par Sylvain MOREAU
Chef du Service de la donnée et des études statistiques (SDES), Commissariat général du Développement durable (CGDD), ministère de la Transition écologique et solidaire

Conséquence du développement d'Internet, les gigantesques fichiers de données produits par l'activité des géants du secteur et issus, en général, des modes de consommation des ménages et des entreprises, représentent des bases d'étude d'une extrême richesse. Leur usage dans le cadre de la statistique publique n'est pas fondamentalement différent du travail que celle-ci a l'habitude de réaliser à partir de fichiers administratifs, si ce n'est que ces données permettent, souvent, de disposer d'une richesse d'informations géographiques et temporelles qui permet d'élaborer des indicateurs plus nombreux et plus détaillés. Suite à la mise en œuvre de la loi Énergie, les acteurs de la statistique publique ont un rôle essentiel à jouer en matière de mise à disposition des informations transmises par les producteurs et les distributeurs. Les données déjà disponibles affichent un niveau de détail géographique jamais atteint jusqu'ici. Les évolutions d'ores et déjà prévues auront un impact important sur les systèmes d'information statistiques portant sur l'énergie et sur les modes de travail de la statistique publique.

Dans le domaine des sciences de l'information, l'expression *Big Data* (ou données massives) a pris ces dernières années une importance croissante. En premier lieu, c'est l'augmentation importante du volume des données produites par les géants de l'Internet, mais aussi par certaines disciplines scientifiques (la génomique et l'astronomie, en particulier) qui a favorisé ce phénomène. Les progrès intervenus dans les techniques de stockage et dans le traitement de données volumineuses de natures variées (notamment en format texte ou image), souvent produites en flux continu (d'où la dénomination des « trois V » (pour Volume, Variété et Vélocité) qui est souvent citée et mise en avant) sont à l'origine de gigantesques gisements de données. Leur existence même a permis la mise en œuvre d'outils et de méthodes spécifiques (par exemple, de certains algorithmes de recommandation pour des sites de vente en ligne), mais elle a aussi donné la possibilité de disposer d'une information extrêmement détaillée, permettant de répondre à des besoins d'information fine, notamment au niveau géographique, qui n'était, tout simplement, pas satisfaits jusqu'ici.

Au final, le *Big Data* recouvre des réalités très différentes : à la fois la mise à disposition de sources de données qui n'existaient pas jusqu'ici et l'apparition de méthodes de gestion et d'analyse de ces données, dont le format peut être non standard.

La statistique publique est naturellement concernée par l'émergence de ces nouvelles sources. Certains instituts nationaux de statistiques (INS) (Istat et CBS, en particulier) ont développé assez tôt une stratégie explicite d'utilisation des *Big Data,* se lançant dans des expérimentations. Plusieurs groupes de travail ont été créés au niveau international (en particulier par Eurostat et l'Unece, mais aussi par l'ONU) pour définir une position commune de la statistique publique sur l'exploitation des nouvelles sources de données, définir des initiatives prioritaires et mutualiser les investissements. À l'Institut national des statistiques et des études économiques (INSEE), le groupe « nouvelles sources » s'est interrogé, dans le cadre des réflexions menées autour d'INSEE 2025, sur le positionnement de la statistique publique sur ces questions.

Dès 2011, l'INSEE (à l'instar de certains instituts statistiques à l'étranger) a lancé un projet intitulé « données de caisse », dont l'objectif était d'intégrer les données issues de la grande distribution dans l'indice des prix à la consommation. Ce projet a permis de mettre en lumière certaines questions à instruire pour pouvoir exploiter ce type de données.

La première de ces questions est juridique : comment pérenniser l'accès de la statistique publique à ces données ? Quand les premiers travaux ont été lancés, il n'existait pas de cadre juridique garantissant cet accès. L'intégration d'un article spécifique dans le projet de loi pour une République numérique a depuis permis d'apporter une réponse à cette question.

La deuxième question est d'ordre technique : il faut garantir la gestion des flux et le stockage de ces données de très grands volumes, ce qui nécessite une infrastructure adaptée. Pour ce faire, l'INSEE a développé une plateforme de stockage dédiée à ce type d'usage.

Les particularités des *Big Data*

Quand on parle de *Big Data,* on pense, en premier lieu, aux données de gestion détenues par des opérateurs privés, dont certaines sont considérées comme étant susceptibles d'être utilisées par la statistique publique (facturations de la grande distribution déjà exploitées par l'INSEE, compteurs électriques intelligents, facturations de la téléphonie mobile…). Les problématiques posées par leur utilisation sont, par certains aspects, assez proches de celles des données administratives, que la statistique publique exploite depuis longtemps : il s'agit de la qualité et de la représentativité des données ; de l'adéquation entre les données dont on dispose et le phénomène que l'on veut mesurer ; de la pérennité de la source et du manque de maîtrise sur les évolutions des formats et des contenus ; et, enfin, de la taille en elle-même de ces gisements de données.

Il s'agit là de problématiques familières à la statistique publique sur lesquelles l'INSEE et les autres services statistiques ministériels (SSM) ont développé des expertises depuis déjà plusieurs décennies.

Ces données massives présentent néanmoins un certain nombre de spécificités que l'on ne retrouve pas chez les sources administratives :

- En premier lieu, leur accessibilité : propriétés d'opérateurs privés, elles font souvent l'objet d'une valorisation économique par ces derniers. Le vote de la loi numérique a donné quelques facilités à l'INSEE et aux SSM pour leur permettre d'accéder à ces gisements de données, mais cet accès est très encadré. Il s'agit de répondre à des besoins d'enquêtes statistiques obligatoires, et ce, uniquement s'il est démontré que ce mode de collecte est adapté aux besoins de l'enquête et qu'il présente, par rapport à d'autres modes de collecte, des avantages en termes de coût pour le service statistique public ou les personnes enquêtées, et/ou de qualité des données produites. Enfin, les renseignements extraits des bases de données ne peuvent être utilisés à d'autres fins que la réalisation de l'enquête statistique spécifique pour laquelle l'accès auxdites bases a été donné.
- La nature même de ces données, qui les rend souvent peu maniables. Pour un certain nombre d'entre elles, elles sont étroitement liées aux comportements des ménages, puisqu'elles en sont directement issues, et leur existence même peut être très corrélée au phénomène que l'on veut observer, qui est de ce fait difficile à quantifier. Elles peuvent donc servir utilement pour qualifier certains modes de consommation, mais elles ne sont que de peu d'utilité pour qualifier et quantifier le pourcentage des ménages touchés par ces phénomènes. Ainsi, il peut être facile en ayant accès aux bases de gestion des sites de covoiturage de disposer de données concernant l'évolution de ce phénomène, les parcours effectués, voire sur certains profils d'utilisateur. En revanche, ce type de données ne donnera pas d'information sur le poids relatif du covoiturage dans les modes de transports globaux ni sur les motivations des déplacements effectués.
- Le caractère sensible de ces données : il s'agit souvent de données personnelles susceptibles de mettre en lumière des comportements individuels, elles sont donc *a priori* sensibles. La possibilité de croiser de manière quasi universelle des volumes considérables de données pour faire émerger de nouveaux services est d'ores et déjà largement utilisée à des fins de profilage, notamment en matière commerciale. La Commission nationale de l'informatique et des libertés (CNIL), très sensible à ces problématiques, a fait évoluer sa doctrine et ses modes de travail de façon à permettre une bonne protection de l'individu, sans pour autant mettre de frein à l'innovation.

L'intérêt du travail statistique sur les *Big Data*

Un des premiers intérêts que la statistique publique peut trouver à travailler sur ce type de gisement de données réside dans l'amélioration et l'enrichissement de la production statistique actuelle.

Il est ainsi souvent mis en avant la possibilité de réduire les délais de publication de certains indicateurs, ce qui constitue un enjeu important pour la statistique publique. L'utilisation de données produites en continu et immédiatement accessibles semble *a priori* un moyen naturel pour réduire les délais de production et de mise à disposition de certains indicateurs : les compteurs intelligents pourraient ainsi produire des estimations plus fréquentes de la consommation d'énergie (électricité, gaz…).

De son côté, Orange promet, à travers son offre commerciale « FluxVision », de convertir, chaque minute, 4 millions de données mobiles en indicateurs statistiques pour mesurer la fréquentation d'une zone géographique et les déplacements des populations.

La masse des données disponibles peut également permettre de produire des indicateurs présentant un niveau de granularité plus fin (sur des sous-catégories ou des sous-populations : les données de caisse, par exemple, peuvent permettre de produire de manière plus systématique qu'actuellement des prix moyens par produit, ou des indices de prix régionaux), ou une plus grande précision, sans pour autant alourdir la charge de l'enquête (par exemple, estimation des temps de transport par mode à partir des données de téléphonie mobile). Elle permet en général de disposer d'informations plus complètes, voire exhaustives, et donc de mettre potentiellement à disposition une information beaucoup plus détaillée, notamment au niveau géographique, et donc plus pertinente pour des acteurs locaux.

Photo © Stéphane Audras/REA

Smart Electric Lyon, une expérimentation à grande échelle d'une gamme de produits et de services nouveaux compatibles avec des « *smart grids* ».

« Les avancées technologiques permettent déjà ou vont bientôt permettre de disposer d'informations sur les consommations en temps réel, et ce avec un niveau de détail inégalé jusqu'ici. »

Ces nouvelles sources de données pourraient en outre réduire la charge d'enquêtes, faire baisser les coûts de collecte et générer ainsi des économies.

Sur ce dernier point, la situation est bien évidemment plus complexe. La réduction de la charge d'enquêtes doit être mise en regard du coût des investissements nécessaires pour assurer le traitement de données massives, qui, par nature, sont moins faciles à appréhender. Ainsi, l'expérimentation de l'utilisation des données de caisse (dans le cas français) a nécessité le développement d'une infrastructure informatique adaptée au traitement de flux de données très volumineuses et l'achat de catalogues permettant de traiter ces données. Un bilan récemment publié par *Statistics Norway* montre que l'exploitation de ces données a effectivement réduit la charge d'enquêtes, mais qu'elle a aussi conduit à une augmentation du nombre des ressources internes allouées à la production de l'indice des prix.

Cette augmentation s'explique non seulement par l'intégration pour la production de cet indice de données plus complexes, mais également par un « effet qualité » se traduisant par la production d'indicateurs potentiellement plus nombreux. Comme cela a été évoqué plus haut, l'information disponible est structurellement plus riche et plus complète, elle est donc susceptible de valorisations beaucoup plus importantes.

Pour un certain nombre de problématiques, l'utilisation de ce type de gisement de données pourrait, à terme, modifier profondément le système d'observation. Compléter l'exploitation de ces sources très détaillées par quelques enquêtes de cadrage pourrait être une solution alternative à des enquêtes régulières, et cela pourrait permettre de continuer à mettre à disposition la même information que celle que nous produisons actuellement, tout en élargissant le spectre des données disponibles, notamment à l'attention des acteurs locaux.

L'énergie, une bonne candidate pour ce type de travail

L'énergie est un domaine dans lequel l'exploitation des *Big Data* est sans doute riche de potentialités. Certaines de celles-ci sont déjà mises en œuvre.

L'énergie est en effet un secteur d'activité qui a subi de profondes transformations, ces dernières années. Les acteurs ont beaucoup changé, leur nombre s'est accru et ils se sont diversifiés. Les avancées technologiques permettent déjà ou vont bientôt permettre de disposer d'informations sur les consommations en temps réel, et ce avec un niveau de

détail inégalé jusqu'ici. Enfin, le principe d'une ouverture des données collectées par les gestionnaires de réseaux d'énergie qu'a posé la loi relative à la transition énergétique pour la croissance verte a, en l'espace d'une année seulement, complètement changé la donne.

En effet, cette loi fait obligation aux transporteurs et aux gestionnaires de réseau de communiquer à l'administration les données liées à la consommation totale annuelle d'électricité, de gaz, d'hydrocarbures et de chaleur, par secteur d'activité et à l'échelle des quartiers. Celle-ci doit les rendre publiques, une fois qu'auront été traitées les questions de secret, celle principalement de la protection des données personnelles. L'objectif premier est de faciliter l'exercice par les collectivités locales de leurs nouvelles compétences en matière d'énergie (élaboration des Schémas régionaux d'aménagement, de développement durable et d'égalité des territoires – SRADETT – et des plans Climat-air-énergie territoriaux – PCAET), notamment) et d'inciter les consommateurs à maîtriser leurs consommations. C'est le Service statistique du ministère de la Transition écologique et solidaire, qui, étant le point focal de cette mise à disposition, a la responsabilité de cette rediffusion de données en direction des acteurs locaux.

Ces données sont disponibles sur le site du Service statistique du ministère précité depuis la fin de l'année 2016 sous un format ouvert aisément réutilisable et exploitable. À un horizon de deux ans, ces données seront disponibles à l'échelle des bâtiments.

Cette offre de données en *open data* constitue *de facto* un véritable levier en termes d'usages nouveaux et de services énergétiques.

À terme, elle va complètement bouleverser les processus de production des données statistiques relatives à l'énergie.

Côté statistiques, on suit actuellement la consommation finale d'énergie des secteurs pris dans leur ensemble, mais aussi par grand secteur économique (agriculture, industrie, transport, résidentiel, tertiaire). La décomposition sectorielle de la consommation d'énergie se fonde sur différentes sources statistiques collectées et traitées par le SSM du ministère chargé de l'Environnement (le SDES, Service de la donnée et des études statistiques), qui les enrichit à partir d'enquêtes réalisées par l'INSEE ou d'autres services statistiques ministériels, notamment celui du ministère de l'Agriculture. Il dispose aussi, s'agissant de la répartition par usages et de la consommation d'énergie dans le résidentiel et le tertiaire (selon la nature du logement), des données du CEREN (le Centre d'études et de recherches économiques sur l'énergie).

L'ensemble de ces résultats est établi principalement au niveau national. Des travaux de régionalisation ont également été réalisés dans le passé. Pour certains secteurs économiques, l'élaboration de statistiques régionales peut s'appuyer sur des enquêtes *ad hoc*. Pour d'autres, il s'agit d'estimations ou de calculs par soldes (c'est notamment le cas du résidentiel pour lequel aucune source statistique présentant un degré de précision suffisant n'est disponible au niveau régional). L'expérience prouve que ces résultats peuvent être fragiles et soulever pas mal de questions. Or, une des fortes demandes adressées aux services statistiques concerne justement la mise à disposition de données fiables au niveau local.

L'accès aux données de consommation à des niveaux géographiques extrêmement fins permet d'apporter une première réponse. Et, typiquement, les difficultés (que nous avons pointées au début de cet article) inhérentes à certaines données *Big Data* ne concernent pas ces types de données. En effet, elles sont exhaustives, ne présentent pas de biais de sélection et sont accessibles de par la loi.

Ceci étant, elles ne permettent pas de répondre à l'intégralité des questions qui sont posées. Il n'est, par exemple, pas possible de connaître les usages. Mais l'on peut imaginer, quand les données seront effectivement disponibles au niveau des bâtiments, qu'il sera alors possible de disposer, par recoupement avec d'autres sources (par exemple, le fichier SIREN), d'une description très fine des consommations d'énergie par secteur d'activité. Cela, en revanche, n'exonérera pas de la nécessité de disposer d'autres outils statistiques (enquêtes, sources administratives, modélisations) qui permettront d'enrichir ces données et de compléter l'information déjà disponible.

Le déploiement des « compteurs intelligents » à la fois pour l'électricité et le gaz pourrait aussi, à terme, ouvrir des perspectives intéressantes pour le suivi de la consommation d'énergie, pouvant même aller jusqu'aux usages, qui pourraient être modélisés pour les ménages. Dans le cas de l'industrie, le niveau de nomenclature pourrait être affiné. Un suivi infra-annuel serait possible à des niveaux géographiques extrêmement fins. Enfin, l'étude de la précarité énergétique pourrait également être approfondie.

Le positionnement des SSM dans la mise à disposition de ces données est un peu différent du positionnement traditionnel d'un service statistique national. En effet, la statistique publique met à disposition des données qui sont considérées comme des données de référence, dont la fiabilité est avérée. Or, dans le cadre de la mise à disposition des données de consommation d'énergie, le temps nécessaire entre la collecte desdites données auprès des opérateurs et leur mise à disposition n'a pas permis de fiabiliser ces données. Seules les erreurs les plus flagrantes, telles que celles constatées dans les unités, ont fait l'objet d'un traitement spécifique. Il est quasi certain qu'il existe des erreurs de saisie : indétectables au niveau national, elles peuvent néanmoins avoir un impact important à l'échelle territoriale. Cela nécessiterait sans doute une qualification de la donnée diffusée, au minimum à un niveau simple (de type donnée « examinée », ou « non examinée »).

Mais si ce type de travail est amené se généraliser, il sera alors sans doute nécessaire de réfléchir à la mise en place de réseaux permettant de fiabiliser les données, avec, par exemple, une décentralisation du travail de correction, dans un premier temps, à l'échelon des directions régionales de l'environnement, de l'aménagement et du logement (DREAL), puis, éventuellement, auprès d'autres acteurs locaux, avec un retour formalisé des données auprès des producteurs de celles-ci, ce qui nécessitera d'inventer des règles permettant un travail collaboratif autour de ces données.