PRÉSENTATION DU NUMÉRO

Chaque numéro du Courrier des statistiques obéit à sa propre logique. Certes, le hasard des propositions d'articles ne permet pas d'aboutir spontanément à une homogénéité de thèmes, et la logique du numéro n'apparaît qu'*a posteriori*. Le numéro N9 est caractérisé par le choix d'articles portant sur des sujets parfois ardus, et inhabituels pour la revue. Nous allons donc tenter ici de faciliter leur compréhension.

Une fois n'est pas coutume, commençons par la fin, avec les trois derniers articles, qui portent sur des sujets liés, et qu'on pourrait à tort considérer comme plus adaptés à une revue d'informatique que de statistique publique. Ils sont en réalité essentiels dans un « monde de *data* », où statisticiennes et statisticiens vont de plus en plus puiser des données externes, administratives par exemple, pour leur propre usage[1] .

Ainsi l'article n°5, écrit par ***Alexis Dondon et Pierre Lamarche***, porte sur les formats de données. En première approche, on pourrait considérer qu'il s'agit d'un sujet annexe, d'une dimension purement opératoire, secondaire pour l'usage statistique. Il n'en est rien : qu'ils soient imposés à l'utilisateur ou au contraire délibérément choisis, les formats ont des propriétés, présentent des limites, des opportunités. Les auteurs expliquent qu'il n'existe pas de format idéal, mais au contraire que chaque format répond à une gamme de besoins, de contraintes. Ils présentent le format Parquet, moins connu, plus récent, adapté à de très gros volumes de données.

Ces données, on se les procure « ailleurs », dans des sources administratives. Il y aurait beaucoup à dire sur cette notion de source, mais partons de ce qui existe et voyons comment on élabore des statistiques à partir de cela. Il faut intégrer les données puis les transformer pour les rendre aptes à être digérées par un processus de production statistique. C'est cette phase méconnue de transformation que décrivent ***Franck Cotton et Olivier Haag*** dans l'article n° 6. Ils en décomposent les étapes, du recodage à la pseudonymisation en passant par le contrôle, du renommage au filtrage en passant par la caractérisation des unités statistiques. Ils insistent également sur la nécessité d'en faire un traitement automatisé et réplicable, un véritable *pipeline* piloté par les métadonnées. La gestion des formats est ici un des aspects importants du processus de transformation et de contrôle.

Vous avez dit contrôle, transformation ? ***Bertrand Dubrulle, Olivier Rosec et Christian Sureau*** (Cnav) s'y intéressent dans l'article n°7, mais dans un contexte très différent : les échanges massifs de données au sein de la protection sociale, par exemple pour l'alimentation de référentiels, comme le Répertoire de gestion de carrière unique (RGCU), ou pour les déclarations administratives. Pour maîtriser les flux de données transmis, et dans une optique de traitements automatisés, la structure attendue des données et les règles qu'elles respectent doivent être très clairement définies : c'est ce qu'on appelle une *norme d'échange*. Dans un contexte où cette structure peut évoluer fréquemment en raison des changements de réglementation, la Cnav a mis au point un outil (Saturne) qui permet de décrire formellement une norme et de générer automatiquement, sur cette base, toute la documentation associée et les outils de contrôle. Une telle démarche est particulièrement pertinente pour assurer la qualité des données, sujet essentiel pour les statisticiens.

1 Voir le dossier « Le statisticien et les sources administratives » du numéro N1 de décembre 2018.

Remontons d'un cran dans le sommaire, avec deux articles (les numéros 3 et 4) portant sur la question de la confidentialité des données et la façon de gérer efficacement cette confidentialité.

L'article n° 3 de ***Patrick Redor*** fournit le cadre d'analyse, en posant la confidentialité des données comme un enjeu majeur de la statistique publique, en raison des risques encourus en cas de violation de cette confidentialité. Mais pour l'activité statistique, les éléments d'identification des personnes sont le plus souvent indispensables, et on ne peut se borner à les enlever. Il faut donc des mesures de protection et un cadre juridique, qui s'est enrichi dans le temps : loi de 1951, loi Informatique et Libertés, loi pour une République numérique, règlement général sur la protection des données (RGPD), *Data Act*. Les règles du secret statistique, si elles ne figurent pas dans la loi, se révèlent subtiles dans leur application, avec notamment le « secret secondaire ». Tout ceci s'inscrit dans un contexte évolutif, où la demande de données ne cesse de croître, et l'on peut parfois se demander si confidentialité et *open data* ne sont pas deux injonctions contradictoires. Promouvoir un large accès aux données peut avoir pour conséquence paradoxale de… restreindre l'accès à certaines statistiques.

Pour assurer la protection des données confidentielles tout en profitant de la richesse des données provenant de sources différentes, il existe une possibilité : s'appuyer sur un « code statistique non signifiant » (CSNS). Ainsi, dans chaque fichier à apparier, on enlèvera les éléments d'identification, en ne conservant que ce code. Celui-ci servira de pivot pour l'appariement, tout en ne permettant pas de remonter à l'individu. Comme l'expliquent ***Yves-Laurent Bénichou, Lionel Espinasse et Séverine Gilles*** dans l'article n°4, le CSNS, plus qu'un code, est un véritable « service » rendu par l'Insee à l'ensemble du service statistique public. Il peut s'appliquer à des NIR (numéro de sécurité sociale), auquel cas l'opération est un pur chiffrement, ou à des traits d'identité (nom, prénom, date et lieu de naissance). Dans le second cas, un algorithme préalable d'identification, i.e. la détermination du NIR à partir des traits, est nécessaire. L'article explicite les différentes étapes de cet algorithme et la mesure de la qualité de l'identification. Celle-ci est indispensable, car la procédure CSNS est entièrement automatisée : disposant des niveaux de qualité, l'utilisateur décidera des seuils à appliquer.

Notre cheminement dans le sommaire nous conduit à l'article n°2, portant sur un sujet sophistiqué, multiforme et en même temps novateur : les comptes nationaux redistribués. ***Mathias André, Jean-Marc Germain et Michaël Sicsic*** en expliquent de façon synthétique les tenants et les aboutissants, ce qui constitue un véritable défi pédagogique. Il s'agit, dans un premier temps, de se replacer dans le cadre « classique » de la comptabilité nationale (certes dans une vision simplifiée)… pour aussitôt faire un pas de côté en posant des questions nouvelles et en inventant pour cela un nouveau cadre, centré sur les ménages. De façon directe ou (très) indirecte, les ménages sont destinataires finaux des revenus et des transferts des autres secteurs institutionnels, par exemple à travers les services rendus par les administrations publiques (santé, éducation, etc.). Ce cadre permet ainsi de construire un revenu « avant transferts » et « après transferts ». On s'intéresse alors aux mécanismes

de redistribution (élargie, complémentaire à l'approche usuelle monétaire), selon le niveau de vie, par cohorte d'âge, par catégorie socio-professionnelle, réconciliant ainsi comptabilité nationale et statistiques sociales. Les auteurs explicitent la genèse de cette démarche, les sources utilisées, la méthode de calcul et en détaillent les principales hypothèses. L'article propose quelques enseignements à tirer et trace des perspectives opérationnelles pour cette méthode internationalement reconnue et promise à un bel avenir.

Ce n'est pas d'avenir, mais de passé dont il est question dans l'article n° 1. ***Gaël de Peretti et Béatrice Touchelay*** nous racontent une histoire, celle de la statistique publique dans les 40 années qui ont suivi la création de l'Insee, sous l'angle de son insertion dans le débat social et politique. Dans une première période, dite de construction, on va poser des bases qui vont structurer le fonctionnement de l'institut : les enquêtes « ménages », avec le souci d'étudier les conditions de vie des ménages, le cadre de la comptabilité nationale, la loi de 1951, la coordination statistique, dans un contexte où l'intérêt de la statistique publique est loin d'être acquis. Apparaissent très tôt des controverses, par exemple sur l'indice des prix. Mais c'est encore un auditoire limité, une institution au service d'un petit nombre de décideurs. Dans la deuxième période, dite de consolidation, de nouveaux publics, de nouvelles ouvertures apparaissent dès les années 60 : la création des observatoires économiques régionaux, la création du conseil national de la statistique qui deviendra plus tard le conseil national de l'information statistique (Cnis). C'est aussi l'ouverture vers le grand public, avec la création d'un département de la diffusion, et une reconnaissance croissante *via* plusieurs succès éditoriaux. Entre-temps, on est passé de la mécanographie à l'informatique.

La suite de cette aventure à lire dans un prochain numéro de la revue, probablement en 2024 !

Pascal Rivière
Directeur de la collection, Insee