UTILISER LES DÉCLARATIONS ADMINISTRATIVES À DES FINS STATISTIQUES

*Pascal Rivière**

Pour compléter les données d'enquête ou pour s'y substituer, l'utilisation des données administratives par les statisticiens se généralise. Pourtant, étrangement, le sujet fait l'objet de très peu d'investigations méthodologiques générales. Le cas particulier des données issues de déclarations administratives est intéressant parce que le processus de production associé présente des similitudes avec l'appareillage statistique. Avec un acteur particulier : l'organisme chargé de la gestion des flux de données, souvent différent des administrations utilisatrices. Celui-ci épargne du travail aux statisticiens et le facilite sur certains plans… même si ces derniers perdent le pouvoir sur une partie des opérations. In fine ***le fichier de données élaboré résulte d'une co-construction statistico-administrative. Pour optimiser celle-ci, il faut souligner l'importance d'une relation suivie et formalisée avec l'organisme concentrateur, ce qui donne au statisticien plus de possibilités d'intervention qu'il n'imagine.***

RICHESSE DES SOURCES ADMINISTRATIVES, FAIBLESSE DE LA MÉTHODOLOGIE

Dans l'arsenal du statisticien, les sources administratives occupent une place croissante, et ce pour des raisons variées. Elles permettent de compléter les données d'enquête en les enrichissant, dès lors que l'appariement est possible : la mise en place future du « NIR haché » offrira de ce point de vue des perspectives tout à fait nouvelles. Il arrive aussi que ces données se substituent aux enquêtes, dans un contexte conjoint d'explosion des possibilités en matière de données (Elbaum, 2018), mais aussi plus classiquement de baisse généralisée des taux de réponse[1] et de restrictions budgétaires. Enfin elles peuvent constituer une source nouvelle, différente des domaines explorés habituellement par l'appareil statistique.

Pourtant, les statisticiens qui utilisent ces données savent qu'il y a loin de la coupe aux lèvres, entre autres parce que les processus de production statistique et administratif diffèrent. Or s'il existe un vaste corpus méthodologique dédié aux statistiques d'enquêtes[2] (voir par exemple Lyberg *et alii*, 2012), on ne trouve rien d'équivalent pour celles qui découlent de sources administratives[3]. Certes, il existe une littérature abondante sur la manière de traiter tel ou tel gisement de données particulier, les problèmes rencontrés

* Chef de l'Inspection générale, Insee.

1. « *Plunging response rates to household surveys worry policymakers* », *The Economist*, 24 mai 2018.
2. Traitant de tirage d'échantillon, de calage sur marge, de calcul de variance, de traitement de la non-réponse, etc.
3. Alain Desrosières écrivait à leur sujet en 2004 : « *aucune théorie des erreurs de portée générale n'existe encore* ».

et les solutions apportées[4]. Mais il est difficile de trouver un article transversal donnant un cadre méthodologique général et opérationnel – à l'exception peut-être de Rouppert (2005) – comme le fait la théorie des sondages par exemple. Dans la littérature anglo-saxonne, on peut citer essentiellement un article de Hand (2018) qui effectue un tour d'horizon des problématiques soulevées par les données administratives.

Cette absence d'armature conceptuelle conduit parfois à une forme de bricolage dans l'organisation de la production, à un déficit de méthode. Et en parallèle, on ne s'interroge pas vraiment sur ce que les données administratives ont de spécifique, en les abordant, sans autre forme de procès, comme un tout indifférencié immédiatement disponible, un fichier de données à traiter.

Le but du présent article est d'interroger les spécificités d'un ensemble particulier de données administratives : celles qui sont issues de déclarations administratives. On observe en effet des proximités bienvenues avec la production statistique ***(encadré)*** et dans le même temps des limitations qu'il est essentiel de connaître.

UN STATUT ET DES USAGES SPÉCIFIQUES

Lorsqu'on associe l'adjectif « administrative » à « donnée », que veut-on dire, au fond ? Le sujet serait vaste à explorer. On va se limiter ici à trois spécificités : statut, usage, mouvement.

Le terme « administrative » revient d'abord à qualifier les données par leur origine, et plus que cela par leur statut. Il signifie qu'elles viennent de la sphère administrative en général : les administrations, au sens de la comptabilité nationale, plus généralement les organismes de droit public, et ce qui relève des devoirs de l'État par ailleurs. Il en découle souvent un degré important de centralisation, tout du moins en France, ce qui facilite la tâche des statisticiens.

Le caractère administratif des données nous informe également sur leur usage[5] : elles s'inscrivent dans les processus de gestion d'un ou plusieurs organismes (que nous appellerons « administrations » par la suite), qui ont leur propre agenda et leurs propres objectifs. Contrairement aux données statistiques dont le seul but est d'informer, leurs consœurs administratives doivent leur existence aux actions qu'elles permettent de déclencher. Elles ne sont pas neutres. Par exemple, la donnée individuelle « montant d'une retraite », pour un individu donné, permet de calculer le montant effectif à payer, puis de procéder au paiement, donc d'agir concrètement dans le monde réel.

L'usage peut aller au-delà de l'administration concernée. Ainsi, les informations du Répertoire National Commun de la Protection Sociale (RNCPS) peuvent permettre de repérer des incohérences entre prestations sociales de diverses natures ; la déclaration sociale nominative est utilisée par la DGFIP, les institutions de prévoyance, les organismes de retraite, l'Acoss, pour leurs propres usages de gestion ; les données de SIRENE servent de référence, de preuve pour les entreprises, elles sont utilisées par les chambres de commerce et d'industrie ou par les greffes des tribunaux de commerce.

4. Cf. tous les articles de la session 20 (« Appariements – fichiers administratifs ») des Journées de méthodologie statistique 2018.
5. Desrosières (2004) : « *Une source administrative est issue d'une institution dont la finalité n'est pas de produire une telle information, mais dont les activités de gestion impliquent la tenue, selon des règles générales, de fichiers ou de registres individuels, dont l'agrégation n'est qu'un sous-produit* ».

Encadré. Le processus de production statistique standard d'une enquête

Utilisé comme référence par l'ensemble des Instituts nationaux de statistiques (INS), le *Generic Statistical Business Process Model* (GSBPM)* explicite un découpage standard en activités à partir de 8 étapes principales : définition des besoins, conception, élaboration, collecte, traitement, analyse, diffusion, évaluation. Le GSBPM évoque aussi le traitement des sources administratives, mais on présente ici plus spécifiquement le processus fondé sur une enquête, de façon extrêmement simplifiée, dans les grands principes.

Commençons par les bases. Afin de mener à bien une enquête, le statisticien public a *grosso modo* besoin de trois éléments pour commencer : un cadre légal, la liste des données qu'il veut recueillir, et la liste des entités auprès desquelles les obtenir.

- En France, la liste des enquêtes est fixée chaque année par arrêté du Ministre de l'économie et des finances, après avis d'opportunité du CNIS et avis de conformité du comité du Label, qui permettent d'officialiser l'enquête et de formaliser notamment l'obligation de réponse.
- La détermination des données d'intérêt nécessite d'en avoir explicité le sens, de s'être mis d'accord sur les besoins d'information, sur une manière pertinente de questionner, et sur la possibilité même de les obtenir. De proche en proche, et via des comités d'utilisateurs notamment, on arrive ainsi à un questionnaire, puis à un support de collecte (questionnaire papier, questionnaire électronique pour l'enquêteur ou sur le net…), et on associe à ce dernier des contrôles.
- Définir les entités à interroger, c'est d'abord expliciter le champ, ce qui revient à définir un ensemble en intention : par exemple, les entreprises du secteur du commerce ayant exercé une activité en 2018, en France métropolitaine ; ou bien : les ménages habitant en région frontalière. On construit ensuite une liste en extension, qui va être la base de sondage (par exemple, l'échantillon-maître, ou toute base de sondage issue de Sirene), puis on tire l'échantillon, qui correspondra à l'ensemble des entités à interroger concrètement.

Disposant des trois outils que sont le cadre juridique, le support de collecte et l'échantillon, le processus de collecte peut commencer. Il s'effectue pendant une période bien déterminée, le service statistique prenant contact avec les unités de collecte. Selon les enquêtes, les modes de collecte peuvent être différents (face-à-face, téléphone, internet) et éventuellement être combinés.

Les données collectées sont soumises à des contrôles automatiques : on recherche ainsi les incohérences, ce qui permet de repérer des données ou combinaisons de données douteuses, par exemple une personne qui déclarerait simultanément être retraitée et avoir 20 ans. Les vérifications individuelles réalisées par la suite sont en lien avec ces incohérences potentielles repérées automatiquement. Elles peuvent nécessiter une interaction avec les enquêtés : en direct lors de la collecte pour les enquêteurs, *a posteriori* par les gestionnaires, via les rappels téléphoniques ou les relances dans le cas des enquêtes auprès des entreprises. Notons que le processus de collecte statistique assume l'existence de non-réponses, soit des non-réponses totales, soit des non-réponses partielles.

Il s'ensuit une phase de traitement et d'analyse : gestion de la non-réponse (imputation, repondération), contrôles supplémentaires (macro contrôles, notamment), tabulation, calculs d'évolution, etc. Puis les publications et études, donc la phase de diffusion et enfin l'évaluation.

*Voir sur le site d'Eurostat : https://ec.europa.eu/eurostat/

LA CRISTALLISATION DES DONNÉES, CLÉ DE LA STATISTIQUE ADMINISTRATIVE

Le fait que les données administratives découlent d'un processus de gestion a aussi une conséquence opérationnelle : elles se trouvent dans des bases de données de gestion, toujours susceptibles d'être mises à jour. Ce sont en quelque sorte des données vivantes, qui peuvent être modifiées en fonction d'événements, internes ou externes. Ils peuvent se produire n'importe quand et certains sont totalement exogènes à l'administration, imprévisibles.

Ceci n'est pas sans conséquence pour le statisticien, qui a l'habitude, avec les enquêtes, de récupérer un jeu de données fixées dans le temps et qui se rapportent à des unités statistiques précisément définies. Certes, les valeurs de ces données peuvent changer en fonction des vérifications, redressements… mais fondamentalement on veut parler de la même donnée. Le caractère sans cesse mouvant des bases de données administratives ne peut donc lui convenir : il ne peut travailler que sur des données figées, qui vont se présenter sous forme de fichier. Il s'agira ainsi nécessairement d'une photo à un instant t, mais aussi d'une sélection des variables et des unités qui l'intéressent. Il faut passer d'un monde en mouvement à une cristallisation de ces données sur une période.

Attention, la question de la temporalité n'est pas un sujet simple. Ainsi il faut distinguer au minimum deux niveaux de dates : la date d'obtention de l'information d'une part, et la date de référence de celle-ci d'autre part. Par exemple, on récupère en mars 2018 (date d'obtention) l'effectif salarié d'une entreprise au 31 décembre 2017 (date de référence). Les choses se compliquent lorsqu'on constate que la date de référence peut devenir une période de référence (ex : l'année 2017), et que celle-ci peut nécessiter tout un calcul, toute une reconstitution des données : par exemple, l'effectif moyen en équivalent temps plein de l'entreprise sur l'année 2017.

LA NOTION DE DÉCLARATION ADMINISTRATIVE

Les données administratives peuvent provenir de plusieurs origines. Elles sont souvent issues de processus de gestion interne à l'administration concernée[6]. Mais elles peuvent aussi découler de déclarations administratives.

De quoi s'agit-il ? Une déclaration administrative, c'est l'obligation qui est faite à un certain nombre d'entités (individus, entreprises, organismes publics) de fournir des informations respectant une certaine forme, selon certaines modalités (internet, papier) et temporalités. Par exemple, les différentes déclarations d'impôts (sur le revenu, foncier…) relèvent de dispositifs très précisément documentés, et ils sont obligatoires, sur une périodicité et une plage de temps déterminées (l'impôt sur le revenu est annuel), pour les personnes ou entreprises assujetties à ces impôts. Dans le monde des entreprises, la déclaration sociale nominative est mensuelle (exceptées les déclarations événementielles), et les règles qu'elle doit respecter sont consignées dans un cahier technique versionné et accessible en ligne[7], car la normalisation des déclarations est essentielle.

L'existence d'une forme d'obligation pour les déclarations administratives, souvent supportée par des textes législatifs et réglementaires et adossée à une documentation très normée, est la bienvenue pour le statisticien. Cette obligation se traduit en effet par un pouvoir coercitif, pouvant prendre diverses formes, comme celle d'engager des poursuites… ce qui réduit, sans non plus l'annihiler, le risque de non-déclaration.

6. Voir article de C. Chambaz sur les statistiques de la justice dans ce même numéro.
7. Voir article de C. Renne sur l'utilisation de la DSN dans ce même numéro.

Attention cependant, la déclaration administrative elle-même n'est pas, *stricto sensu*, une source administrative : ce n'est pas un fichier, c'est un flux. Et la façon de construire un fichier figé de données à partir de ce flux est d'ailleurs un sujet en soi. En particulier, élaborer ce fichier ne se limite pas à empiler les déclarations. Ainsi, pour la DSN, l'entreprise peut faire des déclarations rectificatives, des déclarations sur des périodes antérieures, mais aussi des déclarations événementielles, qui conduisent à modifier des données. De ce point de vue, le processus déclaratif diffère du processus d'enquête car rien n'empêche d'avoir pendant la période « de collecte » des modifications à l'initiative du déclarant, liées ici à la vie de l'entreprise. Un travail de reconstitution et même de consolidation, parfois complexe, est donc nécessaire pour bâtir, à partir de l'ensemble des déclarations, une source administrative. Dans certains cas, la situation est encore compliquée par l'utilisation de plusieurs types de déclarations administratives.

L'ORGANISME CONCENTRATEUR, «HUB» DES DÉCLARATIONS

Cette logique de gestion de flux conduit à effectuer une distinction fondamentale entre deux entités : d'une part l'administration, utilisant les données pour sa propre gestion, d'autre part l'organisme concentrateur. Le concentrateur est celui qui pilote le processus déclaratif dans son ensemble : il réalise le système d'information nécessaire, le documente, anime les instances de gouvernance et groupes de travail, organise le contact avec les déclarants, gère les nomenclatures, assure la communication et effectue le suivi du processus. Il concentre l'ensemble des flux, entrants en provenance des déclarants, sortants en direction des administrations utilisatrices. Mais il ne fait pas nécessairement partie de ces administrations, son rôle de gestionnaire de flux étant distinct.

C'est typiquement le rôle du Gip-MDS[8] avec la plate-forme net-entreprises qui permet d'effectuer des déclarations sociales comme la Déclaration sociale nominative (DSN)[9]. Mais on peut aussi citer l'ATIH[10] avec la plate-forme e-PMSI[11], conçue pour le recueil et l'analyse des informations sur l'hospitalisation, ou le SANDRE[12], qui garantit l'interopérabilité des systèmes d'information relatifs à l'eau (même si dans les deux cas le formalisme déclaratif est moins affirmé que pour la DSN). Gérer de telles infrastructures de recueil d'information de façon industrielle est un métier en soi qui requiert une grande technicité (informatique, maîtrise d'ouvrage, droit), une organisation rodée et réactive, et une gouvernance incluant organismes utilisateurs et contributeurs. Les administrations utilisatrices sont par exemple les organismes fiscaux ou sociaux dans le cas des déclarations sociales, la Caisse Nationale d'Assurance Maladie (Cnam) et les hôpitaux dans le cas des données PMSI, l'Agence Française pour la Biodiversité ou les Agences de l'Eau dans le cas du SANDRE.

Avec la déclaration, nous sommes ainsi dans une logique de flux normés, clairement positionnés dans le temps, liés à une obligation déclarative pour des entités déterminées. Si le processus déclaratif est rigoureusement géré par le concentrateur, on sait ce qui arrive aux données depuis la source primaire unique d'information. En particulier, les contrôles sont réalisés à un endroit et un seul et ils sont documentés.

8. Groupement d'Intérêt Public pour la Modernisation des Déclarations Sociales.
9. Voir article d'E. Humbert-Bottin sur la DSN dans ce même numéro.
10. Agence Technique de l'Information sur l'Hospitalisation (https://www.atih.sante.fr/)
11. https://www.epmsi.atih.sante.fr/ Le PMSI, Programme de Médicalisation des Systèmes d'Information, rendu obligatoire en 1996, vise à définir l'activité des unités du service public hospitalier pour calculer leurs allocations budgétaires.
12. Service d'Administration Nationale des Données et Référentiels sur l'Eau.

DES PROXIMITÉS AVEC LA STATISTIQUE D'ENQUÊTE : LE CHAMP, LES VARIABLES...

L'organisme concentrateur gère un processus qui présente des points communs avec le processus statistique ***(encadré)***, en commençant par le triptyque champ / variables / cadre légal.

Le champ est en principe formellement défini et connu : c'est l'ensemble des personnes, ou entités, assujetties à la déclaration. Il fait l'objet d'une définition en intention, avec un cadre général, mais aussi des exemptions, et des cas particuliers ; par exemple, les entreprises et cabinets « ayant des difficultés à assurer les paramétrages sur la partie des organismes complémentaires » ont été exemptés de DSN en phase 3.

La liste de données à recueillir peut faire l'objet d'âpres discussions entre les parties prenantes, car il s'agit de prendre en compte les besoins de chacun sans multiplier les données à collecter (c'était tout le problème avec les DADS). Ceci, en se fondant sur le droit et en recherchant la mutualisation via un modèle conceptuel de données commun, différence décisive entre DADS et DSN. Les données à recueillir sont définies et documentées, de même que la manière dont les échanges sont normés : on pense au cahier technique de la DSN[13], mais aussi à la documentation très riche du Sandre (dictionnaire de données, modèle de données, scénarios d'échange, règles organisationnelles et techniques des données de référence). Par ailleurs, l'organisme concentrateur élabore, maintient et diffuse les nomenclatures de référence, ce qui est un travail considérable : ainsi, pour l'ATIH, la classification internationale des maladies (CIM), ou la classification commune des actes médicaux (CCAM).

> *« La liste de données à recueillir peut faire l'objet d'âpres discussions entre les parties prenantes, car il s'agit de prendre en compte les besoins de chacun sans multiplier les données à collecter [...]. »*

... ET LE FONCTIONNEMENT DE LA COLLECTE

Reste à collecter les données. Notons d'abord que le support de la déclaration, à savoir un formulaire - type Cerfa - ou quelque chose d'équivalent, ressemble parfois furieusement à un questionnaire, proposé en ligne. La collecte est pour l'essentiel exhaustive sur le champ considéré. Sous cette réserve, l'exhaustivité est obtenue via un suivi rigoureux des non-déclarants, beaucoup plus puissant que le suivi des non-répondants : le pouvoir coercitif résultant de l'obligation de déclaration est sans commune mesure avec celui, limité, des instituts statistiques sur les non-répondants ; chacun connaît les conséquences d'une déclaration d'impôts non transmise ou en retard. On constate pourtant souvent une étrange sous-utilisation (et même non-utilisation) par le statisticien de l'information très riche issue de ce suivi des non-déclarants. C'est le cas par exemple pour la DSN.

À l'instar des questionnaires statistiques, les formulaires administratifs font l'objet de contrôles automatiques. Leur logique est légèrement différente des contrôles statistiques. Ainsi, lorsqu'on déclare une DSN et qu'un contrôle n'est pas satisfait par la déclaration, celle-ci ne « passe » pas, elle est bloquée, c'est comme si elle n'avait pas été envoyée ; le déclarant doit donc en envoyer une nouvelle, jusqu'à ce que sa déclaration « passe ». On trouve des contrôles d'appartenance à une liste prédéfinie (ex : la liste officielle des codes pays, celle des codes postaux), de cohérence formelle entre données (ex : égalité entre l'année de naissance et celle qui figure dans le NIR), de la nature d'une donnée (numérique, alphanumérique, date...), voire de sa structure fine (cf. les règles de forme complexes

13. Disponible sur le site dsn-info : www.dsn-info.fr

des adresses mail). Mais ces vérifications automatiques se limitent à des contrôles « durs », qui doivent absolument être vérifiés : on veille donc à ne pas multiplier les contrôles, car cela bloquerait le processus déclaratif.

EN CONTREPARTIE, UNE PERTE DE MAÎTRISE POUR LE STATISTICIEN

La manière dont la collecte d'informations se prépare et s'organise, si elle présente des similitudes importantes avec le processus statistique (en gros les quatre premières étapes du GSBPM), a cependant un inconvénient majeur : toutes les étapes ainsi décrites sont sous la responsabilité du concentrateur, et échappent donc au service statistique. Champ, données, formulaire, tout cela peut évoluer dans le temps, sans que le statisticien ait le moindre contrôle sur ces aspects. Par exemple si une catégorie de population n'est plus assujettie à tel ou tel impôt et disparaît ainsi du champ, le statisticien n'a aucun pouvoir sur cette décision. Il n'a rien à dire non plus sur la liste des variables collectées ni sur leur définition même : celle-ci fait l'objet de multiples discussions avec les organismes utilisateurs de ces données[14], dans lesquels les services statistiques ne sont pas ceux qui pèsent le plus. Il perd donc la maîtrise des concepts, du cadre de la collecte d'informations.

Il perd aussi la responsabilité de la collecte elle-même, au profit d'un tiers, l'organisme gestionnaire de flux. Les services de production statistique voient au passage disparaître un aspect essentiel de leur travail : le contact direct avec les unités de collecte, et la possibilité de vérifier l'information immédiatement auprès de l'enquêté. Cette interaction avec le terrain, cet ancrage au réel, s'évanouit. Au caractère vivant et dynamique de la collecte se substitue un fichier, projection statique et sans âme d'un processus externalisé.

> *« Plus généralement, le statisticien ne décide pas des modalités d'obtention de ces données : l'organisation du processus déclaratif, les relations avec les déclarants [...], le système d'information dans son ensemble. Enfin et surtout, toute cette organisation peut aussi évoluer dans le temps, en lien avec de nouvelles dispositions législatives, sans qu'il soit lui-même consulté. »*

Plus généralement, le statisticien ne décide pas des modalités d'obtention de ces données : l'organisation du processus déclaratif, les relations avec les déclarants (FAQ, communication), le système d'information dans son ensemble. Enfin et surtout, toute cette organisation peut aussi évoluer dans le temps, en lien avec de nouvelles dispositions législatives, sans qu'il soit lui-même consulté.

DANS LES FAITS, UNE CO-CONSTRUCTION STATISTICO-ADMINISTRATIVE

Les déclarations administratives ne fournissent pas une information « clés en main » aux services statistiques. Il reste au statisticien tout un travail à faire en aval, car les données sont souvent inutilisables telles quelles. Pour aboutir à un fichier de données individuelles propre et adapté à leurs usages, les statisticiens doivent ainsi enclencher un second processus après le process purement administratif. Le résultat est donc issu d'une co-construction entre les deux univers.

14. Cf. les nombreuses structures de gouvernance associées à la DSN, et auparavant aux DADS, faisant intervenir tous types d'organismes de protection sociale, services fiscaux... et statistiques.

Les statisticiens doivent ainsi réaliser plusieurs opérations :

- Des contrôles supplémentaires, qui n'ont pas été effectués côté administratif : on pense aux contrôles de vraisemblance, qui vérifient la plausibilité d'une combinaison de données sur des sujets sensibles sur le plan statistique. Par exemple, on va vérifier la cohérence entre le libellé de profession déclarée et le code PCS, ou vérifier la cohérence de montants de revenus entre eux.
- Des transformations des données : les données de la déclaration administrative sont en quelque sorte des données pivot dont chaque administration fait son miel. Cela vaut aussi pour les services statistiques : pour obtenir les vraies données utiles aux études, il faut parfois calculer de nouvelles données à partir de la déclaration. Ces « transformations » peuvent prendre des formes très simples (ex : calcul de l'âge à partir de la date de naissance, ou de la tranche d'âge à partir de l'âge). Elles peuvent résulter d'un calcul plus complexe (ex : la détermination des périodes d'activité à partir des dates de début et de fin se trouvant dans des déclarations séparées), ou nécessiter des transformations de nomenclatures, dont le niveau de finesse est parfois inutile pour ses usages (ex : on n'a pas toujours besoin de tout le détail de la NAF). Simples ou non, ces transformations doivent être définies rigoureusement pour être appliquées de façon systématique.
- Le passage à de nouvelles unités statistiques, ce qui relève d'un degré de complexité supérieur : par exemple la DSN raisonne sur des unités telles que le salarié, le contrat, alors que le besoin des statisticiens porte plutôt sur le poste de travail. Il faut donc reconstituer les données correspondantes, ce qui demande beaucoup de rigueur.

Ce faisant on effectue une véritable mise en forme et mise aux normes de qualité statistique des données, qui vient se substituer à la mise en forme et aux normes administratives d'origine.

Au fond, l'organisme concentrateur réalise toute une partie du processus de production statistique à la place du service statistique (collecte, interaction avec les déclarants, suivi des non-déclarants, une part importante des contrôles) et ce dernier complète par d'autres contrôles, des redressements, des transformations et une mise aux normes statistiques. Avec un bémol important, la perte de maîtrise : on n'impose rien à l'organisme en question.

LA QUALITÉ DES DONNÉES, UNE QUESTION-CLÉ

Utiliser les données administratives à des fins statistiques requiert que celles-ci soient « de qualité »... L'article de David J. Hand sur le sujet mentionne à plusieurs reprises des « problèmes de qualité » sur les données administratives. Mais de quoi parle-t-on ? Les questions de non-qualité des données en général font l'objet de toute une littérature (par exemple McCallum 2012), mais pour un usage statistique elles se jouent à plusieurs niveaux : conformité de la population au champ souhaité, validité de la sémantique des données (qui peut diverger de ce que souhaite le statisticien), existence de contrôles suffisants (certaines données jugées peu importantes ne sont même pas vérifiées), problèmes de dates, d'unités statistiques...

Dans le cas des données issues de déclarations administratives, le fait que des contrôles aient déjà été effectués dans le processus déclaratif, et qu'ils soient précisément documentés et versionnés, constitue déjà un premier pas, une garantie de premier niveau. Chaque contrôle correspond au fond à une propriété que les données déclarées doivent respecter : appartenir à telle liste, être de nature numérique, être présent si telle donnée est présente... Mais il s'agit là de garanties de forme, de structure, qui n'assurent pas une qualité suffisante.

Car au-delà de la syntaxe, il faut se mettre d'accord sur une sémantique commune, clairement définie. Dans les exemples cités plus haut, comme la DSN ou le système d'information

de l'eau, il existe ainsi, bien avant toute considération sur la normalisation des échanges, un modèle conceptuel de données commun appliquant les standards du langage de modélisation unifié (UML). Il conduit à définir rigoureusement la signification des données, les concepts et liens entre concepts. Sur ces questions, il faut souligner que la Belgique a joué un rôle précurseur, avec la mise en place dès 1990 de la « Banque Carrefour de la Sécurité Sociale », institution publique de sécurité sociale chargée de l'échange de données entre les institutions de sécurité sociale… et exemple typique d'organisme concentrateur. Avec cette Banque Carrefour, véritable source d'inspiration pour la DSN, un modèle conceptuel de données a peu à peu été élaboré, mis en commun (Robben *et alii*, 2006) puis imposé jusque dans la loi qui lui a conféré un caractère opposable. Grâce à cela, tous les flux de données provenant des entreprises se fondent sur une même sémantique, sur les mêmes nomenclatures versionnées régulièrement, le tout étant intégralement documenté et mis en ligne. Ce système, qui était donc très en avance sur son temps, est aujourd'hui toujours opérant et efficace.

HYPOTHÈSE DU MONDE CLOS ET *BACKTRACKING*

La sémantique que nous évoquons ici caractérise un monde qui n'est pas immuable. Ainsi des combinaisons de données qui peuvent être jugées en anomalie à un instant t ne le seront peut-être plus à un instant ultérieur, parce que le monde a changé (exemple : les couples mariés de même sexe, situation impossible dans une certaine période, possible après). Supposer que l'on peut définir la qualité des données par une liste de propriétés formelles, que l'on peut vérifier via des contrôles (ex : vérifier que dans un couple les deux personnes sont de sexe opposé), revient à implicitement à faire « l'hypothèse du monde clos ». Or, comme on vient de le voir, celui-ci évolue inévitablement.

Comment donner un cadre opérationnel à ces évolutions et en tirer une approche opérationnelle et efficace ? La réponse nous vient à nouveau d'Outre-Quiévrain : dans l'ouvrage de référence tiré de sa thèse, Boydens (2000) fournit un éclairage tout à fait original sur le sujet avec la notion de « temporalités étagées » issues de l'œuvre de l'historien Fernand Braudel. Les données sont soumises à trois temporalités : la temporalité juridique, la temporalité des bases de données, et la temporalité du réel. Les trois ne sont pas nécessairement ajustées, synchrones, loin de là : le juridique peut s'adapter avec retard au réel, les structures de bases de données n'ont pas la souplesse nécessaire pour s'adapter automatiquement aux évolutions.

Pour rapprocher ces temporalités, à défaut de les synchroniser, la méthode du *backtracking* (Redman 1996, Boydens 2000, 2018) offre une démarche opérationnelle efficace, que l'auteur a appliquée… à des données administratives, en l'occurrence des données de sécurité sociale.

Cette méthode consiste à identifier, à partir des violations de « business rules » les plus fréquemment observées à la source au sein de la base de données, les causes structurelles et y remédier en remontant les flux d'information vers l'amont, ce qui permet une action beaucoup plus rapide et beaucoup plus fiable puisque la vérification se fait à la source. Quel est le principe ? Les flux de données, qu'ils soient administratifs ou statistiques, font naturellement l'objet de contrôles. Par exemple, si telle prestation ne peut être accordée au-delà d'un certain plafond de revenus, on va contrôler la cohérence entre les données « existence de la prestation » et « revenu ». Et si l'on observe une incohérence entre les deux sur des données, ceci génère une anomalie. Jusque-là, rien de bien original.

L'élément novateur (entre autres) est de considérer les anomalies comme des objets d'intérêt en soi, et de suivre leur évolution dans le temps par type d'anomalie. Et dès que l'un de ces types d'anomalie se met à devenir fréquent, ceci nous informe, potentiellement,

sur l'évolution du réel sous-jacent (par exemple, dans le cas cité, sur l'existence de dérogations significatives au plafond de revenus). L'étape suivante est donc de revenir à la source pour mieux comprendre cela et agir auprès des fournisseurs de données, ou bien faire évoluer les contrôles.

Cette technique, utilisée en Belgique avec les administrations, permet réellement d'améliorer la qualité des données, via ces allers-retours pertinents avec les fournisseurs, avec « un ROI atteignant plus de 50 % en termes de diminution de la part de présomptions d'anomalies à traiter » (Boydens, 2018). Le processus est même allé beaucoup plus loin puisque la démarche de *backtracking* a été officialisée via un arrêté royal[15] .

UNE AUTRE FAÇON DE PENSER LE TRAVAIL DU STATISTICIEN

En conclusion, les déclarations administratives offrent de nombreux avantages pour les statisticiens, notamment en raison d'un processus de production ayant des similarités, mais elles présentent, comme les données administratives en général, un gros inconvénient : la perte de maîtrise par l'appareil statistique, qui n'est plus décisionnaire sur des aspects majeurs (champ, données collectées).

Pour remédier à ce problème ou tout du moins le contenir, il est crucial pour le statisticien de travailler de façon régulière et structurée avec l'organisme concentrateur des flux, bien au-delà de la fourniture de données : dans la conception des données et des contrôles, la mise à jour d'une documentation de référence partagée, ou la gestion des allers-retours avec les fournisseurs de données, dans l'esprit du *backtracking*. C'est là un changement culturel, mais il permet de limiter les incompréhensions entre monde statistique et univers de gestion. Une telle coopération peut enclencher un cycle vertueux favorable à une amélioration suivie et maîtrisée de la qualité des données… et des statistiques qui en sont issues.

L'utilisation de données administratives pour la statistique, vue ici à travers le cas particulier des déclarations donc de la gestion industrialisée de flux, ouvre ainsi des horizons nouveaux pour la statistique, d'autres modalités de travail, de nouveaux métiers, voire d'autres façons de penser notre activité, qui obligent à rompre avec les schémas habituels.

15. Arrêté royal du 2 février 2017 modifiant le chapitre IV de l'arrêté royal du 28 novembre 1969 pris en exécution de la loi du 27 juin 1969 révisant l'arrêté-loi du 28 décembre 1944 concernant la sécurité sociale des travailleurs.

BIBLIOGRAPHIE

Boydens I., *« Informatique, normes et temps »*, Editions Bruylant, 2000.

Boydens I., *« Data Quality & « back tracking »* : depuis les premières expérimentations à la parution d'un Arrêté Royal », *Smals Research*, mai 2018.

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

Elbaum M., *« Les enjeux des nouvelles sources de données »*, Chroniques, n° 16, CNIS, septembre 2018.

Hand D.J., *« Statistical challenges of administrative and transaction data »*, J. R. Statist. Soc. A 181, Part 3, p. 555–605, 2018.

Journées de méthodologie statistique 2018, session 20 *(« Appariements – fichiers administratifs »)* http://jms-insee.fr/programmejms2018/.

Lyberg L., Biemer P., Collins M., De Leeuw E., Dippo C., Schwarz N., et Trewin D., *« Survey Measurement and Process Quality »*, Wiley, 2012.

McCallum E. Q., *« Bad data handbook »*, O'Reilly Media, 2012.

Redman T., *« Data quality for the information age »*, Artech house, 1996.

Robben F., Desterbecq T. et Maes P., *« L'expérience de la Banque-carrefour de la Sécurité sociale en Belgique. »*, revue des politiques sociales et familiales, n° 86 sur le thème de *« La nouvelle administration. L'information numérique au service du citoyen »*, pp. 19-31, 2006.

Rouppert B., *« Modélisation du processus de traitement d'une source administrative à des fins statistiques »*, document de travail SGI, Insee, 2005