SOCIÉTÉ FRANÇAISE DE STATISTIQUE

La statistique dans la cité n° 15 - juin 2019

Lettre bimestrielle du groupe « Statistique et enjeux publics »

Sommaire du n°15 – juin 2019

Éditorial :	- *Les données dans nos sociétés*
Méthodes :	- *La population de la France a été "ajustée"* - *Le brouillage des données dans le recensement de 2020 aux États-Unis*
Vie des institutions :	- *L'Adisp : le portail d'accès de la communauté scientifique aux données de la statistique publique*
Lu pour vous :	- *Le secret statistique*
Agenda :	- *Les Cafés de la statistique : séances passées et à venir*

Éditorial

Les données dans nos sociétés

La notion de vie privée nous paraît aller de soi. Pourtant, historiquement elle ne remonte à guère plus d'un siècle. En revanche, les secrets de fabrication et le « secret des affaires » sont millénaires et ont beaucoup évolué tant dans les cultures que dans le droit. Relative aux personnes ou relative aux organisations, l'information a un statut différent et n'appelle pas la même protection. Un brevet protège un procédé tout en le rendant public. Un État exige de ses administrés des données privées, comme une entreprise de ses clients et employés : est-ce légitime et qui y a accès ? L'organisation en tire des synthèses, qui sont semblables à des statistiques, pour fonder des stratégies confidentielles. Auxquelles de ces données « administratives » un citoyen a-t-il accès ?

Une statistique publique, comme un résultat scientifique, demande que sa justesse soit contrôlable. Cette transparence des méthodes est-elle valable lorsqu'il s'agit du savoir-faire d'un bureau d'étude concurrent d'autres ?

Quant aux données des personnes, que veut encore dire le secret promis par le statisticien lorsque ces mêmes personnes s'étalent sur les réseaux sociaux ? Ou, lorsque les traces de leurs déplacements, achats, activités sont captées de tous côtés, lorsqu'elles deviennent objet de commerce, de traitements à multiples finalités au profit ou à l'encontre des personnes concernées, autant que de toutes sortes de collectivités économiques ou politiques, locales ou étrangères, officielles ou occultes, … ?

Les notions et les normes les mieux établies vacillent. À cet aggiornamento de l'information dans nos sociétés, les statisticiens ont leur partie à jouer, leur mission à tenir, leur positionnement à revoir. Sans abandonner pour autant leurs problématiques traditionnelles : d'abord les revisiter et les actualiser. Cette clarification acquise, ils participent au grand concert que le monde moderne appelle.

Le présent numéro de *« La statistique dans la cité »* signale trois actualités relatives aux données statistiques : 1/ Depuis un siècle, progressivement, les statisticiens ont conçu que le respect dû à la vie privée et au secret des affaires exige la confidentialité des données qu'ils recueillent. Ils se sont dotés de règles, s'en sont vu imposer et parfois ont eux-mêmes demandé la caution de la loi. Vient de paraître l'ouvrage *Le secret statistique*, qui fait le point de cette question, du moins pour la statistique publique. 2/ Les résultats publiés n'épuisant pas le potentiel d'analyse que les données recèlent, doit-on et peut-on les mettre (anonymisées bien sûr) à disposition d'autrui ? L'Adisp participe à cette mise

à disposition pour les chercheurs. 3/ Une diffusion plus large est envisageable, mais l'anonymat ne suffit pas : les données sont parfois réidentifiables. Pour y obvier, divers procédés ont été imaginés, dont le brouillage. Les Pays-Bas étaient parmi les pionniers. Ici nous évoquons les projets du Bureau of the Census des États-Unis.

Pour nous écrire : ***sep@sfds.asso.fr***

Méthodes

La population de la France a été "ajustée"

Une colonne inhabituelle est parue en janvier dernier dans le bilan démographique annuel de la France publié par l'Insee : « ajustement ». Elle a provoqué une certaine émotion chez les utilisateurs. L'Insee calcule chaque année une estimation de la population du pays au 1er janvier. Cette estimation est publiée lorsque le résultat du recensement pour la même année est disponible, et n'est autre que ce résultat, légèrement corrigé pour tenir compte de la différence de date (la date de référence du recensement se situe dans la deuxième quinzaine de janvier).

En temps ordinaire, les estimations au 1er janvier de deux années successives sont comparables. La différence entre elles est la variation de population du pays. Celle-ci est attribuée au « solde naturel » - excédent des naissances sur les décès – et au « solde migratoire » - excédent des entrées sur les sorties.

Mais, parfois, les estimations successives ne sont pas comparables ; c'est alors qu'est introduit « l'ajustement ». Cela se produit quand les questionnaires ont été modifiés, ou bien les conditions de la collecte, au point que cette modification pèse sur le résultat. L'ajustement est alors indispensable : sans lui, la variation de population non significative serait à tort attribuée soit au solde naturel soit au solde migratoire.

Trois fois dans le passé des ajustements ont été nécessaires : en 1968, en 1999, et en 2006 avec l'introduction du recensement rénové. Cette fois-ci, c'est une modification du questionnaire « feuille de logement » qui est en cause. Destinée à mieux repérer les situations de multi-résidences – enfants de parents séparés en résidence alternée, étudiants logés ailleurs pour leurs études, etc. – cette meilleure interrogation a permis d'éviter des doubles comptes.

L'estimation antérieure de la population était donc excessive : les nouveaux résultats incorporent une baisse qui n'est pas réelle, mais qui représente une correction. C'est cette baisse qu'il faut représenter par un terme d'ajustement (négatif). Du fait de la méthodologie quinquennale du recensement, dans laquelle le résultat d'une année repose sur cinq collectes annuelles successives, l'ajustement sera nécessaire jusqu'en 2022. Au terme de cette période, l'Insee estime qu'il devrait représenter 0,7 % de la population de la France.
L'ajustement perturbe les habitudes des utilisateurs d'un chiffre fondamental. Mais c'est un mal nécessaire, et qui rappelle qu'aucune mesure n'est parfaite. L'instrument de mesure est toujours là, et on ne saurait renoncer à l'améliorer.

À consulter : l'Insee rend disponibles sur son site deux notes techniques détaillant l'origine et les modalités de l'ajustement. Voici leurs adresses :
https://www.insee.fr/fr/statistiques/fichier/3692693/Recensement-changement-questionnaire_2018.pdf
https://www.insee.fr/fr/statistiques/fichier/2383177/Recensement-estimation-effet-questionnaire_2018.pdf

Le brouillage des données dans le recensement de 2020 aux États-Unis

La confidentialité différentielle est une technique d'anonymisation pour les fichiers de données individuelles. Elle consiste à introduire un aléa, ou un brouillage, dans les données réelles, pour éviter l'identification d'individus figurant dans le fichier. Lorsque ces données relatives à un grand groupe sont analysées, cet aléa est alors compensé de manière statistique, et les résultats agrégés sur ces données restent pertinents.

L'U.S. Census Bureau a décidé d'utiliser cette méthode pour la diffusion des résultats du prochain recensement de la population, qui se tiendra en 2020. Cette méthode permet en effet de donner des garanties formelles, c'est-à-dire des preuves mathématiques, sur la possibilité de limiter les informations qu'on peut apprendre sur les individus. Cependant sa transposition du monde académique à celui de la statistique publique a été à l'origine de nombreux défis qui n'étaient pas anticipés par les statisticiens du Census Bureau. Ces défis portent sur la nécessité d'avoir le personnel qualifié et l'environnement informatique adéquat, sur la difficulté de répondre à tous les usages de données confidentielles et sur l'inadéquation des processus de brouillage avec les besoins des utilisateurs…

Le Census Bureau a décidé de relever ces défis. Pour réduire les risques de divulgation de données confidentielles, il aurait pu diminuer sensiblement le nombre de tableaux diffusés. Mais il a choisi de crypter ses fichiers au moyen de méthodes issues de la confidentialité différentielle. Et compte même élargir ensuite cette technique à toutes ses publications statistiques.

Pour en savoir plus :
- un chapitre sur la confidentialité différentielle dans un article de Benjamin Nguyen paru dans « Statistique et société, vol. 2, no 4, décembre 2014 » :
http://www.benjamin-nguyen.fr/papers/ss.pdf
- sur les projets de l'U.S. Census bureau :
https://arxiv.org/pdf/1809.02201.pdf

Vie des institutions

L'Adisp : le portail d'accès de la communauté scientifique aux données de la statistique publique

Le service de l'Adisp (Archives de Données Issues de la Statistique Publique) est, depuis près de vingt ans, chargé de la diffusion des données de la statistique publique au sein de la communauté scientifique des chercheurs et étudiants, principalement en sciences humaines et sociales. Depuis 2018, le service est rattaché à l'unité Progedo (PROduction et GEstion de DOnnées) du CNRS, dont l'objet général est de "développer la culture des données". Entre autres actions, Progedo coordonne les acteurs de la diffusion de données statistiques au travers de son département "Quetelet-Progedo Diffusion" (plus connu sous son nom d'origine, le Réseau Quetelet) qui réunit non seulement l'Adisp mais aussi le service des enquêtes de l'Ined et le CDSP (Centre de Données Socio-Politiques).

Au fil des années, l'Adisp a signé des conventions avec les principaux producteurs de la statistique publique et collecte régulièrement leurs données : l'Insee en premier lieu, mais également les services statistiques ministériels, et notamment la Drees (Ministère de la Santé) et la Dares (Ministère du Travail), ainsi que d'autres acteurs publics (Irdes, Cereq, Cerema…). Pour diffusion au monde de la recherche, les producteurs créent des fichiers dits "FPR" (Fichiers Production et Recherche) à destination exclusive de la communauté scientifique. Dans l'univers des données diffusées, ces FPR constituent un niveau intermédiaire entre les données ouvertes, anonymisées et accessibles librement sur les sites des producteurs (insee.fr…) ou via data.gouv.fr, et les données confidentielles accessibles, sous contrôle, via la "bulle sécurisée" du CASD (Centre d'Accès Sécurisé aux Données). Sans être aussi détaillés que ces données confidentielles dont l'accès est payant, les fichiers FPR permettent aux chercheurs d'accéder gratuitement à des données qui sont sensiblement moins agrégées que les données ouvertes et de mener à bien bon nombre de leurs projets de recherche.

Grâce à l'Adisp, le monde de la recherche a ainsi accès à 160 sources statistiques (enquêtes et données administratives) déclinées en plus de 1 300 références : pour n'évoquer que deux sources parmi les plus emblématiques et les plus diffusées, on peut citer les recensements de la population (de 1962 à 2015) et les enquêtes Emploi (depuis 1962). Et ce catalogue multithématique (démographie, éducation et formation, travail et emploi, revenus, conditions de vie, santé, culture, …) s'enrichit de près de 100 références par an…

Au-delà de la collecte et de la diffusion de données, l'équipe du service constitue, pour chaque référence, une documentation détaillée, consultable sur le **site de l'ADISP**, et propose, en libre accès, le téléchargement de documents (questionnaires, dictionnaires des variables, …) permettant d'en explorer le contenu.

En parallèle, l'équipe de l'Adisp est à la disposition des utilisateurs (diffusion.adisp@cnrs.fr) pour toute question ou demande de précision…

Lu pour vous

Le secret statistique

Comment sont protégées les informations que nous fournissons pour l'établissement de statistiques ? Comment peut-on concilier la protection des données individuelles et la mise à disposition d'information très détaillées pour la recherche ?
Ces questions sont au cœur du secret statistique.
Et c'est précisément le sujet du livre *« Le secret statistique »*, de Jean-Pierre Le Gléau, qui vient de paraitre. C'est la première fois qu'un livre est entièrement consacré à ce sujet, pourtant incontournable pour tout statisticien, chercheur ou datascientist. Après avoir exposé un bilan complet des règles et pratiques qui permettent de garantir le secret statistique, l'auteur s'attache à raconter et expliquer les changements des pratiques au fil des évolutions technologiques et législatives. Un développement particulier est ensuite fait sur la question difficile de la diffusion de données individuelles. Enfin quelques exemples pris dans d'autres pays donnent un éclairage sur les techniques retenues ailleurs.

https://laboutique.edpsciences.fr/produit/1074/9782759823420/Le%20secret%20statistique

Agenda

Les Cafés de la statistique : séances passées et à venir
Les derniers « cafés » se sont tenus,
- le 21 mai 2019 sur le thème *« Open data vs Statistique ? »* avec comme invité Lionel Janin (Sous-Directeur chargé de la valorisation et de la stratégie de la donnée, Commissariat Général au Développement Durable)
- le 11 Juin 2019 sur le thème *« Pauvreté des familles et action publique »* avec comme invité Michel Villac, Président du Haut conseil de la famille, de l'enfance et de l'âge

Le premier café de la saison 2019/2020 aura lieu dès le mois d'octobre

Tous nos lecteurs sont invités à proposer des thèmes qui pourraient être retenus pour de futurs « Cafés de la statistique »

Responsable de l'infolettre : Marion Selz, présidente du groupe SEP
Rédacteur en chef : Jean-Pierre Le Gléau
Secrétaire de rédaction : Jean-Louis Bodin
Webmestre : Érik Zolotoukhine