SOCIÉTÉ FRANÇAISE DE STATISTIQUE

La statistique dans la cité n° 18 - février 2020

Lettre bimestrielle du groupe « Statistique et enjeux publics »

Sommaire du n°18 – février 2020

Éditorial

Plusieurs articles de ce numéro sont l'occasion de revenir sur des questionnements ou des événements ayant déjà fait l'objet de cafés de la statistique ou d'annonces dans des précédentes parutions de notre infolettre.

Un premier article fait part de l'élaboration au niveau européen du concept de « grille communale de densité » qui pourra se traduire par une nouvelle délimitation des villes et amène à se pencher sur le concept même d'agglomération urbaine ; les enjeux des zonages avaient été évoqués lors de notre 39e Café en avril 2010. Le deuxième article évoque le comptage des manifestants que l'actualité sociale de ces derniers mois a remis à l'ordre du jour ; il se réfère notamment au travail effectué en 2014 par une commission d'experts indépendants qui avait été présenté pendant le 88e Café en novembre 2015 par l'un de ces experts, Pierre Muller, également auteur d'un article sur ce sujet dans le Vol. 3 No 3 (2015) de la revue de la SFdS Statistique et Société.

Les deux autres articles se réfèrent à deux événements récents. L'un d'eux a pour but de soulever les questions que pose la récente décision du Ministère des solidarités et de la santé de confier à une filiale de Microsoft la gestion de la plateforme des données de santé ; Cette plateforme avait été présentée dans l'infolettre 14 en avril 2019 et pendant le 122e Café en octobre 2019. L'autre article est une « brève » : le sujet de la mesure du sans-abrisme avait déjà été abordé à l'occasion de notre 119e Café en avril 2019.

Ainsi, Cafés et l'infolettre « La statistique dans la cité » se complètent pour essayer d'expliquer comment la statistique peut et doit éclairer les débats de société. L'actualité nationale ou internationale (retraites, coronavirus, …) est bien là pour rappeler ce rôle essentiel de la statistique. Nous invitons vivement nos lecteurs à réagir aux articles proposés et à en proposer d'autres.

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

Méthodes

Délimitation des villes : nouveaux outils

La Commission européenne (DG AGRI, DG REGIO, Centre commun de recherche et Eurostat) a choisi une définition des zones urbaines qui repose sur l'estimation de la population dans des mailles

géométriques fines indépendantes de tout découpage administratif. Ce travail devrait aboutir à une re-délimitation des villes et des zones urbaines. Cela amène à se poser la question : qu'est-ce qu'une ville ? En statistique, un concept de ville doit être opératoire : des informations chiffrées doivent permettre de le cerner sans ambiguïté. La donnée de base est la répartition spatiale de la population : une ville, c'est une « agglomération d'habitants de taille suffisante ». Dès lors, deux questions se posent : qu'est-ce qu'une agglomération ? qu'est-ce qu'une taille suffisante ?

C'est à l'occasion du recensement de 1954 que l'Insee a introduit la définition d'agglomérations de population qui est encore en vigueur aujourd'hui. Il les a définies comme des ensembles de constructions avoisinantes dont aucune ne soit séparée de la plus proche de plus de 200 mètres et contenant au moins cinquante habitants ; les communes dont la majorité de la population appartient à une agglomération d'au moins 2 000 habitants sont appelées urbaines. L'Insee actualise régulièrement la liste des « Villes et agglomérations multi-communales » au sens de cette définition.

Ce concept a le gros inconvénient de s'appuyer sur le découpage administratif des communes, découpage historique qui n'a pas d'équivalent dans la plupart des autres pays européens. Or les outils statistiques modernes permettent de s'en affranchir pour définir l'urbain. On peut désormais, en géo-localisant les immeubles, estimer la population dans des mailles géométriques indépendantes de tout découpage administratif. Ce travail est mené au niveau européen avec des « carreaux » de 1 km de côté. En France, la source utilisée est une base de données géo-localisées issue principalement des fichiers fiscaux de la taxe d'habitation.

Le concept de zone urbaine peut ainsi être rénové en s'affranchissant de toute limite administrative. Le concept « agglomérations d'habitants de taille suffisante » repose sur deux critères : la densité, en nombre d'habitants par kilomètre carré, et la taille, en nombre total d'habitants. La Commission européenne définit « une maille dense »[(1)] comme « un ensemble de carreaux contigus de plus de 300 habitants chacun, comptant plus de 5 000 habitants » ; et, à l'intérieur de ces zones, un « centre urbain » comme « un ensemble de carreaux contigus de plus de 1 500 habitants chacun comptant plus de 50 000 habitants ». Les carreaux n'appartenant pas aux mailles denses sont dits ruraux. On remarquera le relèvement du seuil de taille : 5 000 au lieu de 2 000.

Rien n'empêche ensuite de revenir au découpage administratif et de classer les communes. Si plus de la moitié de la population de la commune vit dans un « centre urbain », la commune est dite « densément peuplée ». Si ce n'est pas le cas, mais que plus de la moitié de sa population vit dans un centre urbain ou une maille dense, la commune est dite « de densité intermédiaire ». Les autres communes sont dites « peu denses ». Le nouvel outil ainsi constitué a été appelé « grille communale de densité », et est de plus en plus utilisé en concurrence avec le découpage « villes et agglos ».

On reviendra dans un prochain numéro sur les définitions statistiques de la ville qui utilisent plus d'informations que la seule localisation de la population, notamment les déplacements domicile-travail.

(1) Le terme employé dans le règlement européen est le mix de français et d'anglais « cluster urbain ». Mais l'Insee a préféré retenir le terme français de « maille dense ».

Compter les manifestants : encore et toujours un défi

Le comptage des manifestants est plus que jamais un enjeu. Combien étaient-ils hier ? Plus ou moins que le mois dernier ? Selon le résultat, on conclura que la mobilisation prend de l'ampleur, ou au contraire s'essouffle. Et ce « diagnostic », à son tour, pourra influencer la suite des événements.

Depuis très longtemps, des comptages sont réalisés par le ministère de l'Intérieur et par les organisateurs des manifestations, souvent des confédérations syndicales. Les résultats, on le sait, divergent « dans les grandes largeurs » ! Il n'est pas rare que les estimations varient de un à dix : un pour l'estimation officielle, dix pour celle des organisateurs.

Face à cette situation, qu'on peut juger affligeante, deux initiatives ont été prises ces dernières années.

En 2014, une commission de réflexion formée de trois experts indépendants a été réunie par le préfet de police de Paris. Elle a validé pour l'essentiel les méthodes de comptage et les résultats de la préfecture de police, à condition de ne pas oublier qu'il ne pouvait s'agir que d'ordres de grandeur.

Fin 2017, quatre-vingts médias se sont associés pour demander au cabinet d'études « Occurrence » de développer une méthodologie afin de procéder à des comptages de façon indépendante tant des organisateurs que des autorités. Les chiffres établis par Occurrence pour les manifestations de 2018 et de 2019 ont été en général proches de ceux publiés par la préfecture de police – en moyenne 15 %

au dessus ; mais très inférieurs à ceux des organisateurs.

Début 2020, un physicien parisien, Bruno Andreotti, a mis en cause la méthodologie du cabinet Occurrence. À la différence de la préfecture de police, qui utilise toujours des comptages « manuels » de fonctionnaires de police chargés d'observer les manifestants, Occurrence a recours à un algorithme opérant sur des enregistrements vidéo de la manifestation, les comptages issus de cet algorithme étant ensuite « redressés » manuellement en tant que de besoin. Selon l'universitaire, interviewé par « Le Monde », la fiabilité de cette technique est douteuse, et il serait nécessaire « de mettre une grosse barre d'erreur d'au moins 30 % sur ces chiffres ». Des améliorations sont envisagées ; mais dans le même temps, de nouvelles formes de manifestations apparaissent…

Le débat technique n'est donc pas clos. Mais est-ce le plus important ? On peut se le demander. Que le nombre de participants à une manifestation ne puisse être connu qu'à 15 % près, voire 30 %, c'est certainement regrettable, mais pas forcément gravissime. En revanche, que des pans entiers de l'opinion puissent accorder foi à des chiffres cinq ou dix fois plus élevés que ceux que d'autres reconnaissent comme probables, voilà qui peut susciter perplexité et inquiétude.

Références :
« Le décompte des manifestants, science fragile » David Larousserie - Le Monde daté du 22 janvier 2020
« Compter le nombre de manifestants sur la voie publique : une problématique statistique, mais aussi et surtout politique » Pierre Muller - Statistique et société volume 3 n°3 décembre 2015

Evénements

Le ministère des solidarités et de la santé a choisi Microsoft pour gérer les données de santé françaises

Le ministère des solidarités et de la santé a choisi Microsoft Azure pour assurer la mise à disposition des données de santé dans le cadre de la plateforme (dite Health Data Hub). Cette décision a donné lieu à des interrogations et des critiques (dossier de Mediapart le 24 novembre 2019, tribune de médecins hospitaliers dans Le Monde du 10 décembre 2019...).

Les questions soulevées sont de plusieurs ordres :

Quels sont les risques et les mesures de protection nécessaires ?
Dé-identifiées, les données du système national des données de santé (SNDS) n'en demeurent pas moins à caractère personnel : en s'en donnant un peu la peine, on peut ré-identifier facilement des personnes. La protection de la confidentialité implique d'empêcher les intrusions de personnes non autorisées ou le vol des données : la base française, probablement la plus grande base de données de santé chaînées au monde, est un trésor national.
Les données doivent être chiffrées lorsqu'elles ne sont pas utilisées par les personnes autorisées et ne doivent pas sortir des « bulles » où elles sont traitées. En revanche, les chercheurs et autres personnes autorisées peuvent « voir » les données sinon elles seraient inutiles : c'est le risque accepté mais c'est aussi la raison pour laquelle l'accès aux données est soumis à autorisation.

La centralisation des données que ce soit de manière permanente ou temporaire ne crée-t-elle pas par elle-même un risque de fuite ?
C'est l'argument des hospitaliers qui veulent conserver la garde de « leurs » données. Il est admis que les données centralisées attirent les tentatives de piratage et que l'impact d'une intrusion y serait plus fort. Inversement les données hospitalières décentralisées sont souvent mal protégées et - surtout - l'utilité des traitements de données dépend des possibilités de les rapprocher d'autres données. En outre les « signaux faibles » exigent pour être identifiés de données nombreuses (effets des médicaments en vie réelle par exemple). Cela étant, il n'est en effet pas obligatoire de tout regrouper de manière pérenne.

Le choix d'un gestionnaire de « cloud » étranger augmente-t-il les risques ?
Un cloud est une grande boite noire où il faut faire confiance aux engagements du gestionnaire. Sachant que ce dernier dispose quelque part des clés de déchiffrement (pour permettre les traitements en clair), le risque de fuite est plus difficile à contrôler qu'avec un ensemble de serveurs physiques identifiés.
Cela étant, le gestionnaire ne pourrait ré-identifier les données et à plus forte raison les communiquer à un tiers qu'en violation de la loi française, du RGPD et de ses engagements. Mais le risque n'est pas

nul, d'autant plus que les États-Unis peuvent s'appuyer dans certaines conditions sur le Cloud Act(2) pour accéder à ces données. D'autre part, des officines étrangères pourraient être tentées de les voler en toute discrétion.

Avantages et inconvénients par rapport au choix naturel qu'aurait été le Centre d'accès sécurisé aux données ?

Le CASD est un groupement d'intérêt public qui réunit l'Insee, le CNRS, le Groupe des écoles nationales d'économie et de statistique, l'École polytechnique et HEC Paris ; il met à la disposition des chercheurs les données de l'Insee, les données fiscales et a des partenariats avec plusieurs ministères ou départements ministériels (dont la Drees, qui pilote aussi l'accès aux données de santé). Il peut ainsi effectuer des appariements entre des données de sources diverses.

Un autre avantage du CASD est qu'il a fait la preuve depuis plusieurs années de sa capacité à rendre service aux chercheurs.

MS Azure est un gestionnaire de cloud généraliste alors que le CASD est spécialisé dans la mise à disposition de données confidentielles. Les deux sont certifiés ISO 27001 et « hébergeur de données de santé ». Le CASD a choisi d'installer chez ses utilisateurs des postes de travail dédiés (SD box) pour l'accès sécurisé aux données. Un poste de travail coupé d'Internet est protégé contre la prise de contrôle par un pirate, mais c'est une contrainte pour les utilisateurs, que le ministère de la santé n'a pas jugé nécessaire ; d'où son choix à ce stade d'un autre prestataire mobilisable sans appel d'offre dans le cadre d'un marché existant.

(2) Le Cloud Act (Clarifying Lawful Overseas Use of Data Act) est une loi fédérale américaine promulguée le 23 mars 2018 qui permet aux forces de l'ordre ou aux agences de renseignement américaines d'obtenir des opérateurs télécom et des fournisseurs de services de Cloud des informations stockées sur leurs serveurs, qu'elles soient situées aux États-Unis ou à l'étranger.

Un recensement des sans-abri à Paris dans la nuit du 30 au 31 janvier 2020

Dans un rapport publié le 30 janvier 2020, la fondation Abbé Pierre note : Le « vvvPlan quinquennal pour le logement d'abord et la lutte contre le sans-abrisme » lancé en septembre 2017 et qui se décline dans vingt-trois villes, métropoles ou départements, se déploie progressivement, mais se heurte à d'évidentes limites. Parmi ces limites figure la difficulté de s'appuyer sur des données récentes et fiables. Or la reconduction de l'enquête « sans domicile » menée par l'Insee en 2012 n'est pas prévue à court terme. Plusieurs grandes communes ont donc pris des initiatives pour compter le nombre de personnes sans hébergement afin d'adapter leur politique en matière d'aide, de domiciliation ou de logement.

La Mairie de Paris a ainsi organisé pour la troisième fois, à l'occasion de la Nuit de la Solidarité du 30 au 31 janvier 2020, un recensement des personnes sans abri. Ce recensement a été mené par des professionnels du social, accompagnés par environ 1 700 bénévoles, qui ont quadrillé Paris (353 secteurs) à la rencontre des plus démunis. Durant cette nuit, on a recensé 3552 personnes sans solution d'hébergement (chiffre comparable à celui de 2019 : 3 641). 2 629 d'entre elles étaient présentes dans les rues de Paris, 365 dans les bois, les parcs et jardins ou talus du périphérique. S'ajoutent également 558 personnes décomptées dans les gares ou stations de métro, au sein de l'AP-HP, de parkings ou dans les halls d'immeubles du bailleur Paris Habitat. 12 % de ces personnes sont des femmes.

Lu pour vous

Le courrier des statistiques

En janvier 1977 paraissait le premier numéro du Courrier des statistiques. Cette aventure s'était interrompue en septembre 2011, après plus de 130 numéros publiés. À l'ère d'internet de l'Open data et de l'Open government, le projet d'origine semblait obsolète. Une étude de l'opportunité menée par l'inspection générale de l'Insee avait montré qu'il fallait revenir sur ce postulat et que nombreux étaient les cadres du service statistique public (SSP), mais aussi des utilisateurs de ce service et des citoyens, qui regrettaient la disparition de la revue et qui estimaient qu'il manquait désormais un chaînon essentiel dans la chaîne d'information produite par le SSP. Notre groupe avait à l'époque fait part du besoin de connaissance citoyenne que cette revue permettait de satisfaire. C'est pourquoi nous avons salué avec enthousiasme la relance de cette revue emblématique en décembre 2018.

Le numéro 3 paru en décembre 2019 consacre plusieurs articles à l'innovation dans la statistique publique : utilisation des données de caisse pour le calcul de l'indice des prix à la consommation,

certification de recherches fondées sur des données confidentielles accessible par le CASD, plateforme de collecte de données sur les entreprises par Internet, présentation de deux nouveaux règlements européens sur les statistiques d'entreprise et les statistiques sociales.

Ce numéro est en libre accès sur le site : **https://www.insee.fr/fr/information/3622502**. il est possible d'en obtenir une version imprimable à partir de ce site.

Annonces

Economie et Statistique a eu 50 ans

a revue Économie et Statistique, créée en 1969 pour publier les travaux de l'Insee, devenue depuis 2017 Économie et Statistique/Economics and Statistics, a fêté son cinquantenaire en 2019 : un numéro spécial paru le 18 décembre 2019 a réuni une série d'articles portant sur les grandes tendances de l'économie française au cours du demi-siècle écoulé en matière de partage de la valeur ajoutée, de croissance et de répartition des revenus, de transformations du marché du travail et de la structure sociale et d'inégalités. D'autres articles de ce numéro spécial évoquent les nouveaux enjeux de l'économie : intelligence artificielle ou encore préoccupations environnementales.
Initialement conçue comme la « revue centrale de l'Insee », elle a évolué au cours de cette période pour s'orienter vers la recherche, en s'ouvrant aux contributeurs extérieurs. C'est bien désormais une revue de niveau académique, publiée par un institut statistique public, éditée simultanément en français et en anglais.

11ème Colloque International francophone sur les sondages

Le 11e Colloque International francophone sur les sondages sera organisé du 14 au 16 octobre 2020 par l'Université libre de Bruxelles(ULB) sur son campus du Solbosch. Ce colloque biennal est organisé depuis 1997 par le groupe « Enquêtes, modèles et applications » de la SFdS. Il permettra de faire le point sur l'état des pratiques et de la recherche dans les divers domaines de la méthodologie des enquêtes et des sondages, de réfléchir au rôle des enquêtes et des sondages dans l'ensemble des méthodes de recueil des données, ainsi qu'à leurs applications dans diverses disciplines telles que la statistique publique, les sciences politiques et sociales, le marketing, la santé ou les sciences de la vie. Pour plus d'informations, on peut consulter son site : **https://sondages2020.sciencesconf.org**

Rencontres de statistiques appliquées de l'Ined : 19 mars 2020

L'Ined organise régulièrement des Rencontres de statistique appliquée. Le thème de la prochaine rencontre (19 mars 2020 de 9 h. 30 à 17 h.) sera : « Données de santé : Enjeux et applications ». La plateforme des données de santé (Health Data Hub), qui était le thème du Café d'octobre 2019 et à laquelle un article est consacré dans ce numéro de l'infolettre est au cœur de ces enjeux. Pour plus d'informations sur cette rencontre :
https://www.ined.fr/fr/actualites/rencontres-scientifiques/seminaires-colloques-ined/rsa-19-mars-2020/

Suite au Café du 14 janvier sur la mesure de la pénibilité au travail

Un article de François Ecalle paru le 13 janvier 2020 sur le site FIPECO (Site d'informations sur les finances publiques): "La pénibilité du travail et l'âge de départ en retraite" nous a été signalé :
https://www.fipeco.fr/pdf/0.05958700%201579078369.pdf

A propos du projet de loi de programmation pluriannuelle de la recherche

La SFdS a signé une tribune parue en page 7 du supplément 'Science et Médecine' du Monde daté du 15 janvier 2020 : « Il faut donner plus de place à l'expertise des chercheurs dans le débat public, la décision politique et l'action collective ». Dans cette tribune, trente sociétés savantes plaident pour que la future loi de programmation pluriannuelle de la recherche publique tienne compte des recommandations de la communauté scientifique.
Les abonnés au 'Monde numérique' peuvent retrouver cette tribune par le lien :
https://www.lemonde.fr/sciences/article/2020/01/14/il-faut-donner-plus-de-place-a-l-expertise-des-chercheurs-dans-le-debat-public-la-decision-politique-et-l-action-collective_6025851_1650684.html

Les Cafés de la statistique

Deux Cafés de la statistique ont été organisés :
• le mardi 14 janvier sur le thème de la mesure de la pénibilité au travail ; notre invité était Thomas Coutrot (Dares – Ministère du travail)
• le mardi 11 février sur la mesure de ce que la planète peut fournir et de ce que l'humanité consomme ; notre invité était David Nerini (maître de conférences à l'université d'Aix-Marseille)

Les prochains cafés auront lieu :
• le mardi 10 mars sur le thème de la **localisation des profits des multinationales** ; notre invité sera François Lequiller (ancien de l'OCDE, retraité de l'Insee)
• le mardi 14 avril sur le thème des **inégalités salariales entre hommes et femmes** ; notre invité sera Nila Ceci-Renaud (Dares) ; il s'agit de la séance qui devait se tenir mardi 10 décembre 2019 et qui a dû être reportée en raison des mouvements sociaux et des grèves de transports.

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