Nations Unies ECE/CES/2017/24

Conseil économique et social

Distr. générale
4 avril 2017
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante-cinquième réunion plénière
Genève, 19-21 juin 2017
Point 4 de l'ordre du jour provisoire
Prochaine génération de statisticiens et de spécialistes de la science des données

Système de production des statistiques 4.0

Note de Statistics Sweden

Résumé

Le présent document examine les qualifications et les compétences nécessaires aux statisticiens compte tenu de l'évolution de la production de statistiques, ainsi que le rôle des statisticiens officiels dans ce contexte. Pour définir ces éléments, il faut analyser le système de production des bureaux nationaux de statistique. Il faut également prendre en considération les utilisateurs et leurs besoins en matière de données, ainsi que la réponse à attendre des bureaux de statistique en termes de production.

Ce document est présenté pour examen au séminaire organisé par la Conférence des statisticiens européens sur le thème de la prochaine génération de statisticiens et de spécialistes de la science des données.

GE.17-05383 (F) 240417 030517
1705383

I. Généralités

1. Les enquêtes fondées sur la théorie de la randomisation ont été et sont encore aujourd'hui une source essentielle de production de statistiques dans les bureaux nationaux de statistique. Un problème majeur dans l'application de cette théorie est la non-réponse. En effet, alors que la théorie présuppose une réponse intégrale, les services nationaux de statistique constatent que, dans la pratique, certains échantillons ne répondent pas. Le problème de l'absence de réponse n'a pas encore été résolu de manière satisfaisante (voir par exemple, Brick, 2013) et compromet aujourd'hui fortement la validité des statistiques issues d'enquêtes par sondage en raison des taux élevés de non-réponse. En partie à cause de ce problème et du coût élevé des enquêtes par sondage, les données massives sont envisagées par les services nationaux de statistique comme une autre source de données possible pour la production de statistiques.

2. L'utilisation des données massives pour la production de statistiques officielles pose de nombreuses difficultés, mais ce type de données présentent aussi nombre de caractéristiques intéressantes (voir par exemple, Daas *et al.*, 2015 ; Japec *et al.*, 2015). Par rapport aux enquêtes par sondage traditionnelles, la collecte de données est considérée comme étant moins coûteuse et plus rapide et peut, dans certains cas, être utilisée pour la production de statistiques en mode quasi-continu. Avec cette méthode, le problème principal n'est plus celui de la collecte des données mais de leur analyse, l'édition des données devenant également une tâche plus complexe (voir par exemple, Tang, 2014).

3. D'évidence, pour qu'un bureau national de statistique puisse passer d'une méthode reposant sur les enquêtes par sondage à un processus utilisant d'autres sources de données, un profil de compétences différent est nécessaire. Dès lors que l'accent est mis non plus sur la collecte des données mais sur leur analyse, les statisticiens doivent posséder les qualifications et les compétences leur permettant d'effectuer ces analyses.

4. Les données massives n'ont pas encore été définies ici. À l'heure actuelle, certains services nationaux de statistique sont fortement tributaires des registres et des données administratives. Les données administratives sont parfois considérées comme des données massives, mais tel n'est pas toujours le cas. Dans le présent texte, l'expression « autres sources de données » sera utilisée en lieu et place de « données massives » et désigne des sources de données autres que les enquêtes par sondage probabilistes et les registres, ces derniers étant définis comme dans l'ouvrage de Wallgren et Wallgren (2014).

5. Pour être en mesure d'examiner les qualifications et les compétences requises des statisticiens et le rôle qu'ils sont appelés à jouer, il faut définir au préalable le système de production des bureaux nationaux de statistique. Il importe également de prendre en compte les utilisateurs et les décisions auxquelles ils sont confrontés, ainsi que les attentes à l'égard d'un service national de statistique en termes de production.

II. Systèmes de production de statistiques 1.0 - 3.0

6. La position solide occupée par la théorie de la randomisation dans les services nationaux de statistique peut s'expliquer par les caractéristiques des statistiques qui en découlent. Les profanes considéreront les statistiques comme étant homogènes et interprétables de la même façon. Ce n'est bien entendu pas vrai et l'interprétation appropriée est faite en prenant en compte la façon dont les statistiques sont obtenues. Une question trop rarement posée parmi les praticiens, et malheureusement aussi dans les milieux universitaires, est celle-ci : « À quoi sert l'inférence statistique ? ». À permettre des estimations par intervalles, qui sont « une spécialité du mode d'induction statistique … », pour citer Chatterjee (2003, p. 30).

7. L'inférence statistique fournit des estimations par intervalles qui, outre une estimation du paramètre à l'étude, renseignent aussi sur l'incertitude des estimations. Les estimations par intervalles sont exprimées sous forme d'intervalles de confiance dans l'interprétation fréquentiste de la probabilité. En vertu de la théorie de la randomisation, les

intervalles de confiance obtenus à partir d'enquêtes par sondage sont objectifs, dans ce sens qu'ils peuvent être interprétés indépendamment des hypothèses formulées. Cette approche est particulièrement intéressante pour les services nationaux de statistique, étant donné le rôle qu'ils jouent au sein de la société. En pratique, des définitions doivent être élaborées concernant les échantillons d'enquête, la population, le cadre, les variables, etc.

8. Les organismes de statistique utilisant les enquêtes par sondage comme principale source de données sont caractérisés ici comme ayant un système de production statistique 2.0. Statistics Sweden a commencé, dans les années 1950, à employer l'échantillonnage aléatoire et a développé le processus de production au cours des décennies qui ont suivi. Le système utilisé avant les années 1950 est appelé « système de production 1.0 ».

9. À la fin du XX^{e} siècle, les données administratives des organismes publics ont commencé à susciter un intérêt, dans l'optique d'un système de production de statistiques fondé sur les registres. Les données administratives offrent plusieurs avantages par rapport aux enquêtes par sondage (voir par exemple, Wallgren et Wallgren, 2014) et, aujourd'hui, la plupart des services nationaux de statistique travaillent à la mise en place d'un système de production de statistiques à partir de registres. Un tel système est bien développé à Statistics Sweden et joue aujourd'hui un rôle aussi important, voire plus important, que les enquêtes, pour la production de statistiques.

10. Même si Statistics Sweden a mis au point à la fois un système statistique basé sur les registres et un système fondé sur des enquêtes par sondage qui fonctionnent bien tous les deux, leurs capacités ne sont pas pleinement exploitées car les statistiques sont produites principalement selon l'une ou l'autre de ces méthodes, et non en les combinant.

11. Dans leur ouvrage, Wallgren et Wallgren (2014, Sec. 15.1.3) préconisent une approche systémique au lieu de la polarisation entre enquêtes à partir de registres et enquêtes par sondage. Les registres peuvent être utilisés pour produire des statistiques, mais ils sont également importants pour la conception des enquêtes par sondage. Les auteurs introduisent la notion d'« enquêtes par sondage basées sur des registres ». Utiliser pleinement les registres disponibles pourrait permettre d'améliorer les estimations des enquêtes par sondage en termes de pertinence et de précision. Ainsi, un système de production 3.0 est défini ici comme un système reposant à la fois sur les données issues de registres et sur celles provenant d'enquêtes par sondage. Les registres et les enquêtes par sondage sont utilisés en parallèle dans la conception des enquêtes.

III. Futur contexte opérationnel

12. Dans son rapport, la Commission sur l'avenir de la Suède (2013, p. 244) déclare : « Nous vivons dans un monde de turbulences, où les frontières s'estompent et les pays deviennent de plus en plus interdépendants, tandis que les mutations s'accélèrent, que, face au flux d'informations sans cesse croissant, il est de plus en plus difficile de déterminer les éléments les plus pertinents et donc les priorités, et que la concurrence au niveau mondial est plus rude que jamais. ». Plus loin, s'agissant des défis démocratiques, la Commission constate (ibid., p. 258) : « Une autre difficulté concerne la complexité croissante des processus décisionnels. Cette situation s'explique principalement par le fait que de nombreux problèmes et défis rencontrés par la Suède débordent le cadre local, régional ou national et revêtent une dimension internationale. Il faudra donc non pas moins, mais davantage de coopération transfrontière pour résoudre les problèmes à l'avenir, ce qui risque de rendre les processus décisionnels encore plus complexes. ».

13. Les statistiques sont produites pour éclairer la prise de décisions et, comme il ressort du constat ci-dessus, les processus décisionnels s'accélèrent dans des contextes qui gagnent en complexité et ont souvent des dimensions internationales. La production de statistiques doit être utile aux processus décisionnels futurs. L'actualité des statistiques et leur fréquence constituent évidemment un aspect et l'on peut également s'attendre à une exigence croissante de comparabilité et de cohérence. De nouvelles statistiques ciblées, destinées à faire la lumière sur de nouveaux problèmes émergents, mais temporaires, sont également à prévoir.

14. Au cours des deux dernières décennies, les données massives sont apparues et ont démontré leur potentiel pour la production de statistiques. On peut s'attendre à voir émerger à l'avenir de nouvelles sources de données et techniques de collecte qui s'ajouteront aux outils existants. Statistics Sweden observe qu'un volume croissant de données administratives est disponible dans différents organismes publics. Cette évolution devrait se poursuivre et déboucher sur de nouveaux modèles d'enquête. C'est évidemment aussi le cas en ce qui concerne le secteur privé, qui offre de nombreuses sources potentiellement utiles pour la production de statistiques. Un problème important à cet égard est l'accès aux données, qui peut soulever des difficultés sur le plan juridique, outre le fait que les propriétaires de données ne sont pas nécessairement désireux de fournir celles-ci.

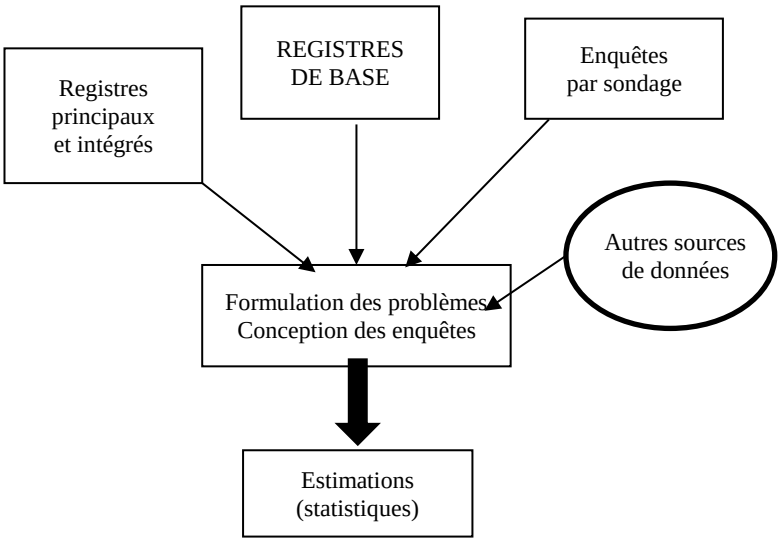
15. La collecte de données dans les enquêtes par sondage fait généralement appel à des questionnaires diffusés sous une forme ou une autre, ou à des entretiens au téléphone ou en face-à-face. Le progrès technologique offrira d'autres moyens de collecte. Il serait aujourd'hui possible d'utiliser les téléphones mobiles dans les enquêtes sur les habitudes de voyage, et on peut concevoir que les transactions effectuées avec les cartes de crédit soient utilisées dans les enquêtes sur le budget des ménages. On peut citer également les connexions Internet des véhicules et le projet Prime Air de la société Amazon. L'avenir apportera davantage de progrès techniques et certains d'entre eux se révéleront utiles pour la collecte de données, permettant non seulement de multiplier les sources de données disponibles, mais aussi d'élargir les options pour la collecte de données spécifiques.

IV. Système de production des statistiques 4.0

16. À la lumière de ce qui précède, l'exigence à laquelle un système de production de statistiques devra satisfaire à l'avenir peut se résumer en un mot : souplesse. Souplesse afin de répondre aux nouvelles demandes, souplesse pour produire des statistiques et en diffuser rapidement les résultats, souplesse pour acquérir, récupérer et traiter les données provenant des nouvelles sources, souplesse pour revoir la conception de l'enquête lorsque de nouvelles sources de données sont disponibles ou que les anciennes sont supprimées, et souplesse dans la communication des résultats. Nous disposons d'au moins trois sources de données : les enquêtes par sondage, les registres et les autres sources de données (par exemple, les données massives).

17. Le tableau schématique d'un système fondé sur ces sources de données est présenté dans la figure 1. Au centre du système se trouvent les registres de base, par exemple les registres de population et ceux des entreprises. Ces registres sont de la plus haute importance pour la validité et l'utilité de l'ensemble de la production statistique. Non seulement ils permettent de suivre l'évolution des unités de population mais ils englobent aussi les domaines et définitions utilisés pour produire des statistiques, ce qui a des incidences sur la pertinence, l'exactitude, la cohérence et la comparabilité des données. À gauche du schéma, on trouve les registres satellites qui conservent des informations sur les unités des registres de base, par exemple le registre de l'emploi et celui des revenus qui sont reliés au registre de la population, et les registres de la taxe sur la valeur ajoutée et du commerce extérieur pour le registre des entreprises. Ces éléments constituent une ressource dans le système et peuvent être utilisés pour produire des statistiques ou à l'appui d'autres enquêtes fondées sur des registres ou des enquêtes par sondage. Certaines statistiques sont élaborées directement à partir des registres de base.

Figure 1
Fonctionnement du système de production des statistiques 4.0



18. Une autre ressource est constituée par les enquêtes par sondage, représentées à droite du schéma. Cet élément renvoie à la fois à la méthode elle-même et aux enquêtes par sondage déjà menées ou à mener. Le cercle à droite représente les autres sources de données, par exemple les données massives.

19. Par rapport aux méthodes traditionnelles de collecte des données dans les enquêtes par sondage, le principe de base de la collecte peut sembler plus simple dans le cas des autres sources de données. Toutefois, dans la pratique, trois grands problèmes sont à prendre en considération. Le premier problème, d'ordre technique, concerne le captage, l'édition et le stockage des données, le deuxième concerne les questions juridiques et éthiques et le troisième concerne le consentement du propriétaire des données à les partager. Une fois captées, les données provenant d'autres sources doivent être traitées et éditées puis analysées selon une méthode appropriée. Ce processus peut impliquer une modélisation statistique beaucoup plus complexe que celle actuellement employée dans les services nationaux de statistique. Il faut aussi combiner et intégrer ces données avec d'autres ensembles de données.

20. Une fois toutes ces ressources en place, il est possible de les utiliser selon différentes combinaisons afin de produire des statistiques dans un domaine donné. On pourra se fonder exclusivement sur les registres ou sur les enquêtes par sondage, ou utiliser l'ensemble de ces ressources.

21. Malheureusement, la théorie et la méthodologie de l'échantillonnage des enquêtes sont souvent considérées par les bureaux nationaux de statistique comme des outils pour étudier les populations qui retiennent leur attention : citoyens, ménages, entreprises, etc. Or, elles constituent des outils généraux, susceptibles d'être utilisés à des fins spécifiques dans d'autres types d'enquête faisant appel à la fois aux registres et à d'autres sources de données.

22. À titre d'exemple, les sondages par interception sont souvent utilisés pour les statistiques touristiques, les visiteurs étrangers étant échantillonnés indirectement sur la base d'un échantillon de points d'interception, comme les aéroports, les ports ou les postes frontière. Ces modèles suivent la procédure normalisée concernant la définition des unités d'étude, la population de ces unités, la conception de l'échantillonnage et la collecte directe de données au moyen, par exemple, d'entretiens. Si le but du sondage est de fournir des données pour les comptes satellites, il suffirait d'estimer le nombre total de visites et les dépenses moyennes par visite. On pourrait aussi concevoir un autre type d'enquête, où l'échantillonnage joue un rôle différent. Par exemple, des données de localisation des téléphones portables pourraient être utilisées en parallèle avec la validation (l'étalonnage) des enquêtes par sondage pour estimer le nombre total de visites. De même, des données fournies par les cartes de crédit pourraient être utilisées en combinaison avec une enquête par échantillonnage pour l'estimation des dépenses moyennes. Dans une telle conception,

les enquêtes par sondage jouent un rôle différent en apportant des éléments d'informations nécessaires que ne fournissent pas les autres sources principales de données.

23. L'élément intéressant dans ce système de production est la dimension « formulation du problème/conception de l'enquête ». Comme le montre l'exemple ci-dessus, en reformulant le problème, on peut concevoir d'autres modalités d'enquête permettant une utilisation plus efficace des ressources disponibles. Une démarche encore plus souhaitable, mais qui ne serait peut-être pas réalisable, consisterait à réexaminer en amont les problèmes rencontrés par les décideurs et motivant l'enquête. La formulation des problèmes et la conception de l'enquête seront déterminantes pour l'utilisation efficace d'autres sources de données.

24. Pour qu'un système de production statistique soit performant et à même de répondre aux exigences et aux défis futurs, le préalable est qu'une coopération étroite s'instaure entre les différents départements du service national de statistique ; spécialistes, méthodologues, informaticiens et autres professionnels doivent œuvrer de concert pour trouver des solutions appropriées à la conception de l'enquête. Combiner les connaissances et les compétences des différents départements est essentiel pour élaborer et tester d'autres modèles d'enquête susceptibles d'être utilisés. Comprendre les besoins des utilisateurs et les contextes permet de définir quelles statistiques il convient de produire ; comprendre les ressources disponibles et comment elles peuvent être utilisées détermine la méthode selon laquelle les statistiques peuvent être produites, sachant qu'il s'agit là d'un processus interactif.

V. Futurs rôles des statisticiens et des spécialistes de la science des données

25. Si l'on garde à l'esprit l'intitulé du séminaire, il apparaît que nombre des nouvelles fonctions et compétences qui caractériseront à l'avenir les services nationaux de statistique ne sont pas abordées ici. Toutefois, elles sont prises en compte et implicitement évoquées dans la description du système de production 4.0.

26. La science des données et ses spécialistes sont des notions nouvelles. On peut trouver plusieurs définitions de ces termes, qui sont souvent associés à la recherche scientifique et à des applications commerciales. Ils renvoient aussi au processus de production de nouvelles connaissances (Boyd et Crawford, 2012). Le rôle principal d'un service national de statistique n'est pas de produire de nouvelles connaissances, au même titre que les théories et les modèles visant à expliquer les phénomènes observés dans le monde réel. Un bureau national de statistique dresse un état des lieux, en produisant des statistiques sur le taux de chômage, le taux d'inflation, le PIB, etc. Il n'a donc pas besoin d'adapter la nouvelle approche destinée à générer des « connaissances » ou des « éclairages », mais il peut utiliser les outils de la science des données et les compétences des spécialistes de cette science pour mettre en évidence des structures cachées dans les grands ensembles de données composites. À partir de là, est élaboré un modèle d'enquête qui produira des statistiques officielles respectant les critères traditionnels de qualité statistique. La production de statistiques officielles pourrait donc constituer un domaine d'application approprié pour la science des données, car la validité des résultats obtenus au moyen du cadre d'échantillonnage fondé sur un modèle ne dépend pas de la « vérité » du modèle.

27. Comme il ressort de la description du système de production 4.0, l'élément clef est la formulation des problèmes et la conception de l'enquête. À ce stade, il sera nécessaire de mutualiser différentes compétences dans une « équipe spéciale », au sein de laquelle statisticiens et spécialistes de la science des données jouent un rôle central. En effet, dans leurs domaines respectifs, ils ont une large connaissance de ce qui peut être accompli avec différentes méthodes et techniques. Une stratégie est donc requise pour garantir l'accès aux compétences appropriées.

28. À l'avenir, les statistiques seront aussi produites au moyen d'enquêtes par sondage, appuyées par des données administratives et des données provenant d'autres sources. D'autres méthodes feront appel à l'échantillonnage pour appuyer les enquêtes fondées sur des données administratives et des données provenant d'autres sources.

29. En outre, dans le futur système de production, les compétences statistiques nécessaires au traitement et à l'exploitation de données administratives seront également indispensables et revêtiront probablement une importance accrue. En effet, de nouvelles sources de données administratives sont constamment générées par les organismes gouvernementaux. D'autres organismes publics disposent également de données qui ne sont pas exploitées actuellement, mais qui pourraient se révéler intéressantes à l'avenir. De même, le traitement des données recueillies auprès d'organisations du secteur privé nécessite des compétences spécifiques, en ce qui concerne les données issues des registres.

30. Enfin, à l'avenir, le rôle des statisticiens et des spécialistes de la science des données portera aussi sur la communication et les moyens de faciliter la tâche des utilisateurs. La méthode traditionnelle consistant à publier des estimations ponctuelles avec les marges d'erreurs correspondantes est acceptable si l'utilisateur a une formation suffisante en théorie statistique. Or, ce n'est généralement pas le cas et les utilisateurs, ne sachant pas quoi faire de la marge d'erreur, n'en tiennent pas compte. Le fait de ne pas interpréter les statistiques selon les exigences traditionnelles d'exactitude est une incitation pour d'autres producteurs à investir le marché, en fournissant des statistiques moins coûteuses mais beaucoup moins fiables et dont l'utilisation peut avoir des effets préjudiciables. Il importera donc de trouver les moyens de communiquer les statistiques de manière à ce que l'utilisateur puisse les interpréter correctement et, dans le même temps, acquérir des connaissances sur les critères d'exactitude qui leur sont propres. Une première mesure consiste tout simplement à remplacer la combinaison d'estimations ponctuelles et de marges d'erreur par un intervalle de confiance, que l'on nommera « estimation par intervalle ».

VI. Références

Boyd, D. et Crawford K. (2012). *Critical Questions for Big Data,* Information, Communication & Society, 15:5, 662-679.

Brick, J.M. (2013). *Unit Nonresponse and Weighting Adjustments: A Critical Review*, Journal of Official Statistics, 29:3, 329-353.

Chatterjee, S.K. (2003). *Statistical Thought - A Perspective and History*, Oxford University Press, New York.

Commission on the Future of Sweden (2013). Future Challenges for Sweden, Cabinet du Premier Ministre, Suède, Ds 2013:19 (English translation).

Daas, P.J.H., Puts, M.J., Buelens, B. et P.A.M. van den Hurk (2015). *Big Data as a Source for Official Statistics*, Journal of Official Statistics, 31:2, 249-262.

Japec, L., Kreuter, F., Berg, M., Biemer, P., et al. (2015). *Big Data in Survey Research: AAPOR Task Force Report*, Public Opinion Quarterly, 79:4, 839-880.

Tang, N. (2014). Big data cleaning. In L. Chen, Y. Jia, T. Sellis, et G. Liu (dirs. publ.) *Web technologies and applications, Lecture notes in computer science*, 8709, 13–24. Cham: Springer International Publishing.

Wallgren et Wallgren (2014). *Register-based Statistics – Statistical Methods for Administrative Data, 2nd Ed*, Wiley, New York.