Nations Unies ECE/CES/2017/25

Conseil économique et social

Distr. générale
11 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

Statisticiens ou spécialistes de la science des données ? L'avenir des statistiques officielles à l'ère des nouvelles technologies et des sources de données modernes[1] [2]

Note établie par le Bureau central des statistiques d'Israël

Résumé

L'essor de la technologie et la disponibilité de mégadonnées font naître des exigences nouvelles en faveur de statistiques officielles plus détaillées, plus précises et plus actuelles. De ce fait, les défis technologiques et méthodologiques qui sous-tendront la production de statistiques officielles dans les années qui viennent auront des répercussions considérables sur le travail des offices nationaux de statistique. Le présent document sera consacré aux questions relatives à la collecte et à la gestion des mégadonnées pour la production de statistiques officielles, à la confidentialité et à la protection des mégadonnées, au développement de l'accessibilité des données sans préjudice au respect de la vie privée et à la confidentialité, à la possibilité d'utiliser les enquêtes en ligne, l'effet modal, les données présentes sur les réseaux sociaux et l'intégration des données administratives aux enquêtes. La question qui se posera ensuite est celle de savoir dans quelle mesure les universités préparent les étudiants au travail au sein d'offices de statistique modernes. Le présent document examine plusieurs de ces questions en s'appuyant sur les expériences nationales et internationales.

Le présent document est soumis pour débat à l'occasion du séminaire sur la nouvelle génération de statisticiens et de spécialistes de la science des données, organisé dans le cadre de la Conférence des statisticiens européens.

[1] Document établi sur la base de l'article intitulé « Methodological Issues and Challenges in the Production of Official Statistics » (« Questions et défis méthodologiques relatifs à la production de statistiques officielles »), rédigé par le professeur Danny Pfeffermann, Bureau central des statistiques d'Israël, à l'occasion de la vingt-quatrième conférence annuelle Morris Hansen et publié dans le *Journal of Survey Statistics and Methodology*, décembre 2015. M. Yoel Finkel, Statisticien associé, a établi la présente version résumée de l'article.

[2] Le présent document a été soumis tardivement faute de ressources.

GE.17-05876 (F) 110517 120517

1705876

I. Introduction

1. L'expression « statistiques officielles » est employée selon une acception très large, mais elle n'a jamais été définie officiellement. Dans le présent document, on entend par statistiques officielles l'ensemble des publications émanant des offices nationaux de statistique, que ces publications reposent sur des enquêtes, des recensements ou des données administratives, ou encore sur une combinaison des trois. Cette description n'en reste toutefois pas moins limitée, car de très nombreuses recherches concernant l'utilisation des mégadonnées en statistique officielle sont actuellement en cours. De façon générale, les mégadonnées ne sont pas le résultat d'enquêtes ; elles sont bien plus volumineuses et dynamiques et peuvent se présenter sous des formats beaucoup plus diversifiés que les données traditionnellement considérées comme administratives. Leur utilisation en statistique officielle représente probablement le défi le plus fascinant que les offices nationaux de statistique aient à relever. Les sections qui suivent seront précisément consacrées à ce défi.

2. Les statistiques officielles sont les statistiques dont on entend le plus souvent parler. Chaque mois paraissent les nouveaux chiffres du chômage, des mesures des revenus et de la pauvreté, des indices de prix, des performances du système éducatif, des statistiques sur la santé et l'environnement et bien d'autres chiffres encore. Pour la plupart des gens, les statistiques se résument aux statistiques officielles. De plus, les responsables s'appuient (ou devraient s'appuyer) sur les statistiques officielles pour mener à bien leur travail de planification et prendre les décisions qui ont des répercussions sur la vie de l'ensemble de la société. Lorsqu'une banque centrale décide de modifier ses taux d'intérêts, elle prend sa décision sur la base de statistiques officielles. C'est aussi le cas des décisions concernant l'affectation des financements publics, la construction de nouvelles écoles, les programmes sociaux, les programmes de santé et même les décisions politiques. Cela étant, on mesure sans peine à quel point il est important de disposer de statistiques officielles actuelles et fiables pour tous les aspects de notre vie. Or, le monde est en constante évolution, des technologies nouvelles et toujours plus perfectionnées sont mises au point et, dans le même temps, les contraintes budgétaires sont toujours plus fortes.

3. Le présent document a pour objet d'examiner certains des défis méthodologiques probablement les plus importants que les producteurs de statistiques officielles ont à relever et, parfois, de proposer des moyens d'y faire face. Dans cette version abrégée soumise à la Conférence des statisticiens européens de 2017, les défis suivants sont examinés :

a) Collecte et traitement des mégadonnées pour la production de statistiques officielles ;

b) Intégration de l'informatique pour la production de statistiques officielles à partir de mégadonnées ;

c) Accessibilité des données, vie privée et confidentialité ;

d) Intégration des statistiques et de l'information géospatiale.

4. Une fois posée l'importance primordiale des statistiques officielles, la question qui vient inévitablement à l'esprit est celle de savoir si les universités préparent les étudiants à travailler au sein des offices nationaux de statistique. Comme on le verra dans la dernière partie du document, cela n'est généralement pas le cas. Il semble même que la situation se soit détériorée au cours de la dernière décennie. Aujourd'hui, seules quelques universités proposent des enseignements de base sur les statistiques officielles, par exemple sur l'échantillonnage. Cette situation est particulièrement inquiétante si l'on sait que les offices nationaux de statistique figurent parmi les plus gros pourvoyeurs d'emplois pour les économistes et les statisticiens.

II. Collecte et gestion des mégadonnées pour la production de statistiques officielles

5. Le terme « mégadonnées » désigne le plus souvent des flux massifs et rapides de données de grande valeur à la fois complexes et de structure, de provenance et de format variables, mais qui comportent aussi des incertitudes intrinsèques qui en amoindrissent la véracité (la définition des 5V des mégadonnées ; il existe également une définition des 7V). Les exemples les plus connus sont les données sur le génome et le cerveau humains, les données liées aux réseaux sociaux et au commerce sur Internet, les relevés d'imagerie satellitaire, les capteurs climatiques, les utilisations des téléphones mobiles, etc. Les immenses difficultés que les scientifiques doivent résoudre pour gérer et analyser ces données font l'objet de nombreuses autres publications[3]. Ces excellents rapports font à peine mention de la production de statistiques officielles, mais les offices nationaux de statistique ne peuvent évidemment pas ignorer les avantages potentiels des mégadonnées du point de vue des statistiques. Diverses initiatives sont d'ores et déjà en cours dans ce sens. Par exemple, en 2014, la Commission de statistique a créé un groupe de travail mondial ayant pour mandat d'« assurer l'orientation stratégique, la direction et la coordination d'un programme mondial sur l'utilisation des mégadonnées en statistique officielle, et de promouvoir l'utilisation concrète des sources de mégadonnées en statistique officielle » (ONU, 2014).

6. Certains aspects importants de l'utilisation possible des mégadonnées en statistique officielle sont présentés ci-après, les problèmes de calcul et de confidentialité étant traités dans les sections suivantes :

a) Type de données : il est important de faire la distinction entre les données obtenues à partir de capteurs, de caméras, de téléphones mobiles ou d'images satellites, qui sont généralement structurées et précises et concernent une population ou une région particulière, et les données issues des réseaux sociaux, du commerce en ligne, de la publicité sur Internet et d'autres supports de même nature, qui sont très diverses et sans structure, paraissent de façon irrégulière et ne portent plus sur une population particulière. Les auteurs du document National Research Council (2013) ont défendu l'idée selon laquelle la structure (ou plutôt l'absence de structure) pouvait évoluer en très peu de temps et que les offices nationaux de statistique devaient être prêts à parer à cette éventualité. En règle générale, les données issues de sources différentes peuvent être codées dans différents formats, arriver à différents moments et avec des degrés de fiabilité différents, voire être définies différemment. Plus inquiétant encore, certaines mégadonnées peuvent brutalement cesser d'exister, imposant de modifier promptement la production des statistiques qui reposent sur cette source. Par exemple, une entreprise de téléphonie mobile peut brusquement se retirer du marché ;

b) Publication : les ensembles de données des deux exemples cités plus haut, de même que bien d'autres ensembles potentiels de mégadonnées, ont pour caractéristique commune que les données sont disponibles pratiquement pour n'importe quelle période. Actuellement, les publications de statistiques officielles sont annuelles ou mensuelles ou portent même sur une seule journée. Trois questions intéressantes se posent :

i) Quels sont les types de statistiques qui doivent être compilés et publiés ? Les publications officielles issues de mégadonnées mesurées en continu doivent-elles paraître principalement sous forme de graphiques et d'images sur Internet ? Les offices nationaux de statistique utilisent déjà des outils performants de visualisation de données, mais la source des données est généralement bien plus simple à utiliser ;

[3] Le récent rapport intitulé « Statistics and Science » (2013) renferme un résumé intéressant. Il résume les exposés présentés et les débats tenus lors d'un atelier spécial consacré à l'avenir de la statistique, qui s'est tenu à Londres (Royaume-Uni) en 2013, avec la participation de 100 invités, dans le cadre des célébrations de l'Année des statistiques. L'Académie des sciences des Etats-Unis a publié un rapport encore plus technique et fouillé, le document intitulé « National Research Council » (2013).

ii) Si on considère que les estimations agrégées (moyennes) continueront de former la base de la planification et des décisions, comment transformer les données d'entrée dynamiques (mesurables en continu), par exemple, en agrégats mensuels ? Les statisticiens devraient-ils sélectionner les données mesurées en continu par échantillonnage ou par d'autres méthodes plus sophistiquées ?

iii) Il semble évident que les échantillonnages aléatoires joueront encore un rôle prépondérant à l'ère des mégadonnées, mais les échantillonnages réalisés à partir de mégadonnées dynamiques seront différents de ceux obtenus à partir de populations finies. Il faudra par conséquent développer de nouveaux algorithmes d'échantillonnage qui, outre qu'ils contribueront à réduire l'espace de stockage, produiront également des ensembles de données manipulables à partir desquels il sera possible de faire fonctionner des algorithmes afin de produire des estimations et de réaliser des modèles, par exemple pour résoudre le problème posé par les échantillons obtenus à partir des réseaux sociaux. La question de savoir si les offices nationaux de statistique devraient utiliser ces données en statistique officielle est un autre problème[4]. L'échantillonnage facilite par ailleurs la protection de la vie privée (voir la section 4.1 plus loin) ;

c) Estimation algorithmique : les méthodes d'échantillonnage traditionnelles établissent une distinction entre estimateurs basés sur le plan de sondage, estimateurs basés sur des modèles et estimateurs assistés de modèles. Dans ce dernier cas, l'estimateur est choisi à partir d'un modèle, mais ses propriétés sont étudiées par échantillonnage à distribution aléatoire. L'utilisation de mégadonnées fait apparaître une nouvelle catégorie d'estimateurs que l'on pourrait appeler estimateurs algorithmiques et qui sont le résultat d'un algorithme de calcul appliqué aux données brutes. Par exemple, il existe en Israël une demande constante de statistiques caractérisant le degré de religiosité de la population juive et d'informations démographiques et socioéconomiques pour les différentes catégories définies dans le cadre de cette caractérisation. Au début de 2006, l'auteur d'un manuscrit non publié[5] a fusionné 12 fichiers administratifs différents avec le registre de la population israélienne. Les fichiers, qui contenaient environ 6 millions d'entrées, avaient été réalisés à partir d'un algorithme hiérarchique complexe qui attribuait une note de religiosité comprise entre 1 et 3 à chacune des personnes figurant dans le registre. Le registre ainsi fusionné couvrait environ 95 % des personnes âgées de 0 à 64 ans ;

d) Mesures d'erreur : les offices nationaux de statistique s'emploient à faire en sorte que les statistiques qu'ils publient s'accompagnent de mesures d'erreur (d'incertitude) exprimées sous forme d'erreurs types ou d'écarts de confiance. Les mégadonnées sont censées ne comporter aucune erreur d'échantillonnage (à moins qu'un échantillonnage soit effectué). Les mesures d'erreur sont-elles encore un problème dans le cas des mégadonnées ? Faudrait-il se concentrer sur les mesures d'écart et de qualité (erreurs de mesure) et non sur la variance ? Comment évaluer les écarts ? Les statisticiens devraient-ils, à un moment donné, procéder à des comparaisons entre ces estimateurs et les estimateurs obtenus au moyen d'enquêtes traditionnelles ?[6] ;

e) Écarts : le risque potentiel d'écarts importants figure parmi les sujets de préoccupation principaux en ce qui concerne l'utilisation des mégadonnées en statistique officielle. Les écarts de couverture et de sélection apparaissent lorsque les données disponibles ne couvrent pas ou ne représentent pas correctement l'ensemble de la population étudiée. Par exemple, il est évident que les prix de vente des maisons affichés sur Internet ne représentent pas l'ensemble des prix de vente sur un mois donné (écart de couverture). Si la collecte de données tend à favoriser les unités les plus importantes (par exemple les plus grandes entreprises), on observe un écart de sélection. Les opinions exprimées sur les réseaux sociaux sont souvent très différentes de celles qui sont défendues

[4] Voir l'argumentaire proposé dans National Research Council (2013).

[5] Portnoi (2007), Bureau central israélien des statistiques (ICBS).

[6] Voir National Research Council (2013) et AAPOR (2015) pour une analyse plus détaillée des possibles erreurs de mesures associées aux mégadonnées.

par le grand public. Un moyen bon marché de remédier à un écart de couverture dont on connaît l'existence consiste à redéfinir la population étudiée. Par exemple, restreindre la population des « maisons en vente » aux « maisons mises en vente sur Internet », mais est-ce bien là la question que l'on souhaite étudier ? Il arrive que l'existence d'écarts de couverture ou de sélection ne soit pas connue ; comme on l'a dit plus haut, une façon de détecter un écart et d'estimer son ampleur consiste à comparer les estimations obtenues à partir des mégadonnées avec les estimations sans écart obtenues par sondage traditionnel (pour autant que ces derniers existent encore) ;

f) Couplage de données : les offices nationaux de statistique, outre qu'ils produisent et publient des agrégats de données nationaux, produisent aussi très souvent des estimations de résolutions bien plus élevées, avec ventilation par âge, sexe, ethnie, lieu de résidence, type d'activité économique, etc. Or, les mégadonnées disponibles ne renferment pas nécessairement l'ensemble de ces informations, et un travail de couplage considérable devrait être accompli si les informations manquantes étaient présentes dans d'autres sources. Cela fait toutefois apparaître une autre limite possible des mégadonnées, à savoir l'absence des identifiants nécessaires au couplage des différents fichiers. Par exemple, les données relatives aux achats en supermarché ne renferment aucune information concernant les acheteurs, ce qui n'est pas le cas des données collectées dans le cadre des enquêtes sur les dépenses des ménages. Les seuls identifiants qui pourraient permettre de relier achats et acheteurs sont les numéros de cartes de crédit, mais on est en droit de se demander si les sociétés qui les délivrent seraient disposées à communiquer les données pertinentes aux offices nationaux de statistique.

7. Il découle de l'énumération qui précède que l'utilisation des mégadonnées en statistique officielle exigera peut-être de nouvelles méthodes, par exemple pour le couplage des données lors de l'extraction de données provenant de sources différentes. Il convient par ailleurs de concevoir de nouvelles méthodes d'édition et d'analyse permettant de traiter suffisamment rapidement et avec suffisamment de précision l'information disponible (qui peut être évolutive), des méthodes de visualisation perfectionnées et des nouvelles méthodes d'estimation et d'évaluation des erreurs. Ce ne sont là que quelques-uns des jalons quant à la façon de procéder.

8. Dans la section qui suit, on s'intéressera à l'architecture informatique et logicielle sans laquelle les offices nationaux de statistique ne seraient pas en mesure d'utiliser les mégadonnées. Dans la section 4, il sera question de la maîtrise de la confidentialité des données et de la protection contre leur divulgation.

9. Le présent document met en lumière les défis considérables que doivent relever informaticiens et statisticiens en ce qui concerne l'utilisation des mégadonnées. D'un autre côté, il serait irresponsable de ne faire aucun cas des avantages potentiels que l'utilisation des mégadonnées en statistique officielle pourrait apporter, que ce soit par l'actualité des données (donnée en temps réel dans certains cas) ou par une polyvalence, une couverture et une précision accrues (en dépit des possibles écarts de couverture). Les mégadonnées ont vocation à demeurer et à devenir de plus en plus importantes. Leur utilisation ne requiert ni cadre d'échantillonnage, ni questionnaire, ni entretiens ni aucun autre des ingrédients nécessaires aux enquêtes par sondage, ce qui pourrait, à longue échéance, permettre de réduire les coûts dans des proportions considérables. Les taux de réponse dans les enquêtes traditionnelles étant en recul constant, il apparaît inévitable de devoir faire appel aux mégadonnées en complément ou en remplacement des sources d'information traditionnelles.

III. Intégration de l'informatique à l'utilisation des mégadonnées en statistique officielle

10. Le volume considérable et la très grande diversité des mégadonnées requièrent des outils informatiques et logiciels modernes et très puissants pour le stockage, le traitement et l'analyse des données, des outils que les offices nationaux de statistique ne possèdent généralement pas actuellement. Lorsque nous stockons des données sur nos ordinateurs portables, nous raisonnons généralement en gigaoctets (environ 109 octets). Or, les

mégadonnées se mesurent généralement en téraoctets (1012 octets), en pétaoctets (1015 octets), voire même en exaoctets (1018 octets) et en yotaoctets (1024 octets). Pour donner un petit aperçu de l'ordre de grandeur de ces chiffres, Eric Schmidt, le Président Directeur général de Google, a déclaré que nous produisions tous les deux jours autant d'information que nous en avions créé depuis l'aube de la civilisation jusqu'à 2003[7]. Cela correspond à environ cinq exaoctets de données. Aux États-Unis, la société de grande distribution Walmart traiterait 1million de transactions par heure et alimenterait une base de données de 2,5 pétaoctets, ce qui représente près de 170 fois le volume des données stockées par la Bibliothèque du Congrès. Il n'est donc pas étonnant que les moyens matériels et logiciels habituels ne suffisent pas à stocker, à traiter et à analyser des volumes de données aussi considérables. De plus, on assiste à un essor prodigieux du nombre de capteurs et de machines produisant des données. Les capteurs météorologiques et les capteurs de pollution, les capteurs de trafic routier, les capteurs mobiles et les systèmes d'imagerie satellitaire figurent parmi les exemples que l'on peut citer et qui, pour certains d'entre eux, l'ont déjà été plus haut.

11. Pour répondre aux besoins de systèmes de calcul aussi performants, l'informatique en nuage semble offrir des solutions intéressantes en ce qui concerne l'accès à des ensembles de données très volumineux et le traitement de ces ensembles, reposant sur des éléments d'infrastructure puissants (installations de stockage et puissance de traitement). Elle permet de créer des machines virtuelles offrant un espace de stockage et une puissance de calcul considérables. L'architecture se constitue d'une myriade de machines virtuelles qui permettent un traitement de l'information segmenté en processus parallèles multiples. Les utilisateurs (les entreprises) peuvent utiliser l'informatique en nuage pour mettre en œuvre leurs services de mégadonnées sans devoir mettre leurs propres infrastructures à contribution. Ce ne sont en fait pas les utilisateurs du nuage qui gèrent l'infrastructure et la plateforme distantes à partir desquelles l'application est lancée. L'informatique en nuage offre tous les logiciels nécessaires et peut aussi permettre de stocker et gérer les voix, ce qui constitue un atout intéressant en statistique officielle.

12. L'informatique en nuage a toutes les chances de jouer un rôle de plus en plus important dans l'accès aux mégadonnées, dans leur traitement et dans leur analyse, même si son développement est encore actuellement freiné par les difficultés rencontrées pour transférer des gros volumes de données. Elle semble malgré tout représenter pour les offices nationaux de statistique une solution attractive en matière d'utilisation des mégadonnées, compte tenu, en particulier, des gains de productivité rendus possibles par le fait que plusieurs utilisateurs peuvent travailler sur les mêmes données simultanément. Pourtant, cette solution pose un problème majeur, à savoir celui de la protection des données. La centralisation permet, en théorie, d'améliorer la protection des données, mais les risques de divulgation augmentent de façon incontestable lorsque les utilisateurs sont nombreux et que les données sont réparties sur un espace et un nombre de machines plus importants. Il faudrait en revanche étudier la possibilité de recourir à des installations en nuage privées (centres de données) qui intégreraient l'ensemble des machines locales utilisées pour stocker les données au même endroit que les outils de traitement dans le cadre d'un système de gestion centralisé. Telle pourrait bien être une des tâches majeures que les offices nationaux de statistique devront accomplir dans un avenir proche.

13. Les lignes qui précèdent ne donnent qu'un aperçu très succinct de la puissance informatique que les mégadonnées réclament et de ce que l'informatique moderne peut offrir. Elles permettent toutefois de comprendre toute l'ampleur des moyens informatiques (espace de stockage, matériel, logiciels, puissance de calcul, outils d'analyse et dispositifs de sécurisation des données) dont les offices nationaux de statistique devront se doter s'ils envisagent d'utiliser les mégadonnées pour produire des statistiques régulières.

14. La plupart des outils informatiques dont les offices nationaux de statistique disposent traditionnellement ne permettent pas de faire face à ces besoins, et il est donc indispensable d'acquérir des infrastructures modernes et adaptées. Il convient tout particulièrement de préciser, à cet égard, que les suites logicielles couramment utilisées par les offices nationaux de statistique, telles que SAS, SPSS et R, renferment déjà plusieurs

[7] http://techcrunch.com/2010/08/04/schmidt-data/.

procédures logicielles applicables aux mégadonnées, mais il faudra que le personnel de tous les niveaux hiérarchiques soit doté de compétences informatiques nouvelles pour pouvoir envisager de produire des statistiques officielles à partir de ces mégadonnées. En tout état de cause, tout ce qui précède sera traité de façon très différente si l'informatique en nuage doit finalement être utilisée. Dans ce cas, l'essentiel de l'effort devra porter sur la confidentialité et la protection des données, et les gouvernements devront engager un travail de supervision et de réglementation. Les coûts à supporter seront peut-être moins élevés, mais, dans un cas comme dans l'autre, il est incontestable que l'utilisation des mégadonnées en statistique officielle demandera aux offices nationaux de statistique un effort considérable dans les années qui viennent. À l'image de ce qui se pratique dans de nombreux pays, le Bureau central des statistiques d'Israël s'est doté d'une équipe spéciale qui étudie la possibilité d'utiliser les ensembles de mégadonnées auxquelles nous serions susceptibles d'avoir accès.

IV. Accessibilité des données, protection de la vie privée et confidentialité

A. Introduction

15. Les chercheurs, les décideurs, les journalistes et le public en général exercent sur les offices nationaux de statistique des pressions constantes, leur demandant de publier des données de haute résolution, voire, si possible, des données individuelles. Ces exigences vont bien sûr à l'encontre de la nécessité de protéger la vie privée et d'assurer la confidentialité des données. Si la confiance à cet égard n'est pas entretenue, aucune enquête ne pourra plus être réalisée. Qui plus est, cette obligation est soulignée dans tout questionnaire écrit et à l'occasion de tout entretien.

16. Cette question comporte deux volets différents : la protection des données contre les intrusions, également connue sous le vocable de « cybersécurité », et la garantie que les données qui sont communiquées à l'extérieur des offices nationaux de statistique ne pourront pas être utilisées pour divulguer des données confidentielles privées.

a) La protection des données contre les intrusions est un problème d'informatique, un problème colossal qui présente de plus en plus de risques et qui, bien entendu, ne se limite pas aux données conservées par les offices nationaux de statistique. Nous devons régulièrement renouveler les ordinateurs et équipements informatiques connexes pour renforcer la protection des données contre les intrusions. Comme toujours avec les problèmes de cette nature, au bout de trois ou quatre ans, les statisticiens peuvent à nouveau s'entendre dire que leurs données ne sont plus protégées et qu'il faut racheter des ordinateurs coûteux pour qu'elles le soient à nouveau.

b) Le second aspect, connu sous le nom de contrôle de la divulgation des statistiques, requiert depuis des décennies une coopération entre statisticiens et informaticiens. Les lignes qui suivent donnent un bref aperçu de certaines méthodes et des mesures de la qualité qui y sont associées, avec un accent particulier sur les nouveaux défis liés à l'utilisation des mégadonnées.

B. Risques de divulgation

17. Traditionnellement, les offices nationaux de statistique publient les statistiques soit sous forme de microdonnées provenant pour la plupart des enquêtes sociales soit sous forme de tableaux de données contenant des dénombrements de fréquences ou des données quantitatives généralement collectées dans le cadre des enquêtes sur les entreprises, telles que les données relatives aux recettes totales. De nombreux travaux de recherche ont été menés pour comprendre comment quantifier les risques de divulgation inhérents à chacune de ces méthodes traditionnelles en appliquant une des méthodes de contrôle de la divulgation des statistiques et comment évaluer l'incidence de la méthode utilisée sur l'utilité des données et veiller à ce que l'information utile aux recherches et à la prise de

décisions soit encore présente dans les données publiées. De toute évidence, plus les données sont protégées contre la divulgation, moins elles sont utiles, et inversement.

18. Une autre forme nouvelle de risque de divulgation est le risque de divulgation inférentielle, c'est-à-dire le risque de donner accès à la connaissance de nouveaux attributs avec une forte probabilité. Par exemple, un modèle de régression à fort pouvoir prédictif peut être à l'origine d'une divulgation inférentielle même pour des individus qui ne sont pas présents dans l'ensemble de données. Un autre exemple est celui de la divulgation par différenciation, c'est-à-dire des situations dans lesquelles des publications multiples sont diffusées à partir de la même source. Ainsi, des tableaux de recensement pourraient être différenciés/manipulés afin de divulguer des données individuelles. Cette forme de divulgation peut être contrôlée de manière plus efficace en limitant les variables et les catégories à un ensemble prédéterminé, ce qui empêche la différenciation des groupes non imbriqués d'individus.

19. Un concept étroitement lié à celui de divulgation inférentielle est le concept de protection de la vie privée fondée sur la différenciation, abondamment étudié par les informaticiens dans le contexte de la protection des données[8]. L'idée de vie privée différentielle consiste à éviter la divulgation inférentielle en empêchant un adversaire de prendre connaissance des attributs d'une unité cible spécifique présente dans la base de données et présentant une probabilité élevée lorsque seule une valeur de la base de données a été modifiée et que l'adversaire dispose d'une information complète sur toutes les autres unités présentes dans la base de données (le pire des scénarios). Cette définition relativement exigeante permet de contrôler la divulgation par différenciation ou issue de modèles hautement prédictifs, ce qui, en comparaison avec l'ancien système des publications imprimées, pose davantage de problèmes en raison de la demande croissante de systèmes de diffusion des statistiques en ligne. La solution proposée par les informaticiens pour garantir la vie privée en se fondant sur la divulgation par différenciation consiste à ajouter du bruit et des perturbations aux résultats des recherches en ligne en fonction de paramétrages spécifiques, mais elle contribue bien entendu à réduire l'utilité des données à des fins d'inférence. En conséquence, les autres moyens de protéger la confidentialité des données font l'objet d'études constantes, et les paragraphes qui suivent en donnent un bref aperçu avant de se conclure sur quelques considérations rapides.

C. Protection des données par l'aménagement d'enclaves de données

20. Depuis une vingtaine d'années, beaucoup d'offices nationaux de statistique dans le monde ont aménagé dans leurs locaux des centres de recherche sécurisés, également appelés enclaves de données. Les enclaves de données sont des environnements sécurisés au sein desquels les chercheurs ont accès à des données confidentielles. Les serveurs sécurisés ne sont connectés ni à des imprimantes ni à Internet et seuls des utilisateurs autorisés y ont accès. Les données ne peuvent pas être retirées des enclaves et les chercheurs sont spécialement formés aux règles de sécurité. Des logiciels de statistique tels que SAS, STATA ou R sont mis à la disposition des chercheurs, et le flux d'information est intégralement contrôlé et surveillé. Toutes les données produites sont vérifiées manuellement avant leur sortie de l'enclave afin d'éliminer les risques de divulgation liés, par exemple, aux petites cellules de dénombrement ou aux courbes résiduelles qui pourraient indiquer l'existence de valeurs incohérentes, ou à l'aide d'estimations par noyaux à faible bande passante.

21. Les enclaves de données présentent pour inconvénients évidents d'obliger les chercheurs à se rendre physiquement sur le site de l'office national de statistique et imposent une charge de travail supplémentaire aux employés qui doivent préparer les fichiers de données requis et administrer le système. Récemment, certains offices nationaux de statistique ont étendu le concept d'enclave de données à l'accès à distance en créant des enclaves virtuelles. Les enclaves virtuelles permettent aux utilisateurs de se connecter à des serveurs sécurisés et d'accéder aux données depuis leur ordinateur personnel, l'ensemble de leurs activités, jusqu'à la saisie au clavier, étant enregistrées et surveillées. Le laboratoire

[8] Voir Dinur et Nissim (2003) et Dwork, et al. (2006) pour davantage de détails.

de données sécurisé doit être approuvé par les organismes et les données produites sont vérifiées à distance par le personnel en charge des questions de confidentialité avant d'être renvoyées aux chercheurs par transfert sécurisé. L'utilisation d'enclaves de données virtuelles requiert évidemment une confiance accrue de la part des offices nationaux de statistique, dont le contrôle est amoindri par rapport à celui qu'ils exercent sur les enclaves de données créées à même leurs locaux.

D. Contrôle de la divulgation de statistiques dans le cas des applications Web

22. Soucieux d'accéder à la demande des décideurs et des chercheurs souhaitant disposer de tableaux de données spécialisées sur mesure et, en particulier, de données de recensement, plusieurs offices nationaux de statistique ont mis au point des serveurs générateurs de tableaux flexibles qui permettent aux utilisateurs de définir et réaliser leurs propres tableaux. Les utilisateurs accèdent aux serveurs via Internet et définissent les tableaux à partir d'un ensemble de variables et de catégories accessibles.

23. Une méthode de contrôle de la divulgation des statistiques peut être appliquée aux données produites sous forme de tableaux selon deux approches différentes : une approche en amont et une approche en aval. La première approche consiste à appliquer la méthode de contrôle de la divulgation des statistiques aux données originales de façon que tous les tableaux générés par la suite puissent être considérés comme susceptibles d'être diffusés en toute sécurité. La seconde approche consiste à produire les données originales d'abord, puis à appliquer la méthode de contrôle de la divulgation des statistiques aux tableaux. Elle repose très largement sur la définition informatique de la notion de protection de la vie privée fondée sur la différenciation, dont il a été question au paragraphe 4.2. Il est également possible d'appliquer les deux approches en combinaison, mais cette façon de procéder peut entraîner une surprotection des données et, par conséquent, une diminution de leur utilité.

24. Pour pouvoir générer les tableaux en toute flexibilité, le serveur doit quantifier le risque de divulgation dans le tableau original, appliquer une méthode de contrôle de la divulgation des statistiques, puis réévaluer le risque de divulgation. Il est évident que ce risque variera selon que les données sous-jacentes proviennent d'un recensement et que les zéros sont réels ou que les données proviennent d'une enquête et que les zéros sont aléatoires. Une fois les tableaux protégés, le serveur devrait aussi calculer l'incidence de l'application de la méthode de contrôle de la divulgation des statistiques sur l'utilité des données en comparant le tableau perturbé et le tableau original. Les mesures basées sur la théorie de l'information peuvent être utilisées pour évaluer le risque de divulgation et l'utilité des données sur un serveur de génération de données en tableau.

25. En règle générale, les serveurs distants permettant de générer des données en tableau sont conçus de façon que de nombreuses règles spéciales de contrôle de la divulgation des statistiques puissent aisément être programmées en amont dans le système de façon à exclure les tableaux qui ne doivent pas être divulgués. Ces règles peuvent notamment consister à limiter le nombre de dimensions du tableau, à définir des seuils minimums de population (tailles moyennes des cellules ou nombre de petites cellules), ou encore à définir des catégories de variables cohérentes et imbriquées pour empêcher la divulgation par différenciation[9].

26. Les méthodes de contrôle de la divulgation de statistiques en amont peuvent notamment comprendre des permutations d'entrées, c'est-à-dire des permutations d'attributs entre deux entrées présentant des similitudes au sein d'un ensemble de variables de contrôle. Les méthodes de contrôle en aval peuvent comprendre la perturbation de cellules, notamment l'arrondissement aléatoire, ou l'utilisation d'une méthode postérieure à la randomisation consistant à perturber les dénombrements de cellules par l'application

[9] Voir Shlomo, Antal, et Elliott (2015) pour en apprendre davantage sur les règles de contrôle de la divulgation de statistiques et les méthodes applicables aux serveurs distants permettant de générer les données en tableau.

d'une matrice de transition de probabilité. La méthode de contrôle de la divulgation des statistiques doit permettre de préserver suffisamment de statistiques, notamment les totaux marginaux, et de maintenir une cohérence entre les mêmes cellules générées dans des tableaux différents pour éviter le risque d'identification de la méthode utilisée.

E. Serveurs d'analyse à distance

27. Un serveur d'analyse à distance est un système en ligne capable d'exécuter dans un environnement sécurisé une requête entrée par un chercheur pour obtenir les données appropriées et de restituer un contenu confidentiel sans qu'il soit nécessaire de recourir à une intervention humaine pour vérifier manuellement les risques de divulgation. Comme dans le cas des générateurs de tableaux flexibles, les requêtes sont saisies à travers une interface à distance et les chercheurs n'ont pas directement accès aux données. Les requêtes peuvent concerner des analyses exploratoires de données, des mesures d'association, des analyses de régression et des expérimentations statistiques. Elles peuvent être exécutées sur les données originales ou sur les données confidentielles et les résultats peuvent être restreints et vérifiés en fonction du niveau de protection requis[10].

F. Données synthétisées

28. Depuis quelques années, les offices nationaux de statistique tendent à produire des microdonnées synthétisées à partir de modèles, ce qui permet de préserver des propriétés statistiques importantes présentes dans les données originales. Les données synthétisées sont stockées sous formes de fichiers en accès public. Elles rencontrent un succès de plus en plus grand, car, comme on l'a vu au paragraphe 4.3, l'accès aux données réelles via des serveurs à distance est susceptible d'être interdit. Depuis quelque temps, la tendance consiste à laisser les chercheurs développer et écrire les logiciels appropriés en s'appuyant sur les données synthétisées, puis adapter les logiciels aux données réelles dans un environnement sécurisé tel qu'une enclave de données.

29. Pour produire des données synthétisées, l'office national de statistique adapte un modèle en fonction des données originales et réalise un échantillonnage des données synthétisée à partir de la distribution postérieure correspondante, un peu à l'image de la théorie des imputations multiples. Plusieurs échantillons de données synthétisées peuvent être extraits pour obtenir des estimateurs de variance valides[11]. Les données synthétisées peuvent être combinées avec une partie des données réelles afin d'obtenir un mélange des deux types de données[12]. Il convient néanmoins d'observer que des ensembles de données partiellement synthétisées peuvent encore présenter des risques de divulgation qu'il convient de contrôler avant de les diffuser. Si les modèles utilisés pour analyser les statistiques sont des sous-modèles du modèle utilisé pour produire les données, l'analyse des données synthétisées devrait produire des inférences valides, à condition, bien sûr, que le modèle original soit « correct ».

30. S'agissant des tableaux de données, des techniques permettent également de construire des tableaux quantitatifs de données synthétisées à partir des statistiques sur les entreprises. L'ajustement contrôlé des tableaux permet de supprimer des cellules et de les remplacer par des valeurs imputées qui préservent certaines propriétés statistiques[13].

31. L'utilisation de données synthétisées doit-elle devenir une façon habituelle de protéger les microdonnées ? Par essence, en statistique, les analystes devraient utiliser des données réelles et non pas des données issues de modèles, même si on pourrait dire aussi

[10] O'Keefe et Good (2008) décrivent l'élaboration de modèles de régression via un serveur d'analyse à distance. O'Keefe et Shlomo (2012) comparent les résultats obtenus à partir des données originales en appliquant deux méthodes de contrôle de la divulgation : les résultats obtenus à partir de microdonnées confidentielles et les données confidentielles obtenues à partir des données originales via un serveur d'analyse à distance.

[11] Voir Reiter (2005) et Abowd et Vilhuber (2008) pour plus de détails et d'analyses.

[12] Voir Little et Liu (2003).

[13] Voir Dandekar et Cox (2002).

que des données perturbées ne sont pas non plus des « données réelles ». Un des problèmes principaux posés par les données synthétisées tient au fait qu'elles dépendent totalement de l'adaptation du modèle en fonction des données originales, une procédure subjective, et que le modèle peut ne pas couvrir toutes les relations entre les variables, particulièrement au sein d'une même sous-population. Il est par ailleurs difficile de reproduire les possibles anomalies présentes dans les données. Ainsi, les données originales peuvent renfermer des observations aberrantes pour des ensembles d'unités particuliers (les entreprises par exemple). Pour ressembler aux données originales, les données synthétisées devraient également présenter des observations aberrantes d'ampleur et de comportements analogues pour des ensembles de données similaires. La confidentialité des données est-elle encore préservée même si les observations aberrantes sont quelque peu modifiées ? Enfin, qu'advient-il des mégadonnées ? Devons-nous produire de nombreux ensembles de mégadonnées synthétiques au risque de multiplier le problème posé par le stockage et le traitement des microdonnées ?

G. Éléments de réflexion

32. Il y a de toute évidence conflit entre la demande des chercheurs, qui souhaitent des données plus détaillées, et la responsabilité des offices nationaux de statistique, qui est de protéger la confidentialité des répondants. De ce conflit naît une incontournable nécessité de mettre en place des méthodes de contrôle de la divulgation des statistiques qui répondent à la double tâche consistant, d'une part, à garantir la confidentialité avec un niveau élevé de probabilité et, d'autre part, à préserver l'utilité des données communiquées aux chercheurs. Cette situation a donné lieu à une coopération étroite entre les informaticiens qui établissent les définitions officielles du risque de divulgation, particulièrement en ce qui concerne la divulgation inférentielle, et les statisticiens qui élaborent des méthodes de contrôle de divulgation qui garantissent la confidentialité. L'utilisation de ces méthodes impose aux chercheurs de devoir effectuer leurs analyses statistiques en présence de données perturbées, ce qui leur demande de parfaire leur connaissance de l'inférence statistique dans un environnement marqué par des erreurs de mesure. Il faut incontestablement continuer à explorer des méthodes de contrôle de la divulgation basées sur la perturbation des données qui permettent de préserver l'utilité des données et, du même coup, d'établir des modèles d'estimation cohérents et fiables.

33. Qu'en est-il des mégadonnées ? Le problème de la diffusion des données, qui se pose avec une acuité accrue, requiert une coopération encore plus resserrée avec les analystes en informatique. D'abord, nous devrons traiter des volumes beaucoup plus importants de données complexes aux dimensions multiples, comportant des variables et des catégories beaucoup plus nombreuses que les enquêtes traditionnelles. Il est notamment possible de se limiter à des échantillons sélectionnés à partir des mégadonnées, non seulement pour réduire la taille des données, mais encore pour préserver leur confidentialité.

34. La diffusion des mégadonnées par les offices nationaux de statistique pose des problèmes de politique publique et d'éthique, qui nécessiteront peut-être l'adoption d'une législation spécifique. Lors d'une enquête ou d'un recensement, la protection de la confidentialité des données fait l'objet d'un engagement clair, ce qui n'est pas le cas en ce qui concerne les mégadonnées recueillies par des capteurs, auprès des opérateurs de téléphonie mobile ou sur les réseaux sociaux. Les entreprises qui collectent les mégadonnées seront-elles tenues de les communiquer aux offices nationaux de statistique ? Le public acceptera-t-il que des données privées soient transférées aux chercheurs et éventuellement diffusées, même après l'application de méthodes de contrôle de la divulgation ? De surcroît, l'accès potentiel à plusieurs ensembles de données et la possibilité d'établir des couplages entre ces ensembles multiplient les risques de violation de la confidentialité. Cela est particulièrement vrai des ensembles de données permettant de suivre les activités humaines, par exemple d'identifier la même personne sur des réseaux sociaux multiples.

V. Intégration des statistiques et de l'information géospatiale

35. L'utilisation des systèmes d'information géographique (SIG) permet d'ajouter une dimension spatiale aux données collectées et offre de ce fait la possibilité de les analyser sous des angles différents[14]. Un exemple bien connu est celui des cartes de la pauvreté établies par la Banque mondiale et d'autres organisations, sur lesquelles chaque région géographique est identifiée par des couleurs différentes en fonction des niveaux de pauvreté. Un autre exemple est celui d'une carte des accidents de la circulation sur laquelle chaque tronçon de la route est coloré en fonction du nombre d'accidents qui s'y produit. Ce type de carte des « points chauds » offre incontestablement pour avantage non seulement la possibilité de visualiser en un seul coup d'œil les zones géographiques (régions, tronçons routiers, etc.) qui appellent une attention particulière, mais encore de mettre en évidence les similitudes géographiques entre zones voisines si elles existent. En d'autres termes, ces cartes transforment des estimations ponctuelles en estimations continues.

36. L'utilisation des SIG offre de nombreux autres avantages :

- Elle permet d'améliorer la conception des enquêtes en délimitant les contours géographiques des strates, des cellules d'échantillonnage, etc. Elle offre également toutes les informations requises aux fins de l'échantillonnage aréolaire ;
- Elle permet de répartir efficacement les quotas d'échantillonnage entre les enquêteurs et d'optimiser les itinéraires que ceux-ci devront emprunter pour se rendre dans les endroits retenus comme échantillons ;
- L'utilisation des SIG permet en outre de surveiller des phénomènes tels que l'évolution chronologique de la situation socioéconomique d'individus ou de ménages résidant dans certains lieux et de les relier à des mesures géographiques telles que les distances par rapport à une grande ville, la possibilité de s'y rendre quotidiennement pour y travailler et les déplacements vers d'autres régions ;
- Les SIG permettent d'améliorer de façon très significative la résolution des données en rendant possibles l'étude et la formation de nouveaux groupes (blocs) qui étaient inconnus auparavant.

37. L'essor rapide des technologies ouvre la voie à la collecte de nouvelles mégadonnées de résolution très élevée et l'utilisation des SIG permettra de relier ces mégadonnées à des niveaux de localisation géographique très détaillés.

38. Jusque-là, il a été observé dans le présent document que le statisticien de l'avenir devrait se former non seulement à la théorie statistique classique, mais encore à l'informatique et à la cybersécurité. Toujours en ce qui concerne les futurs statisticiens, les offices nationaux de statistique devront notamment résoudre trois problèmes de taille : la possibilité de recourir à des sondages par Internet en statistique officielle ; le traitement des effets modaux dans le cadre des enquêtes basées sur un mode mixte ; et l'utilisation combinée des données administratives et des méthodes d'estimation des petites zones. Ces problèmes sont traités en détail dans la version longue du présent document, les méthodes statistiques évoluées qui sont présentées dans l'article ayant été retirées de cette version abrégée.

VI. Les universités préparent-elles les étudiants à travailler au sein des offices nationaux de statistique ?

A. Introduction

39. La question de savoir si les universités préparent les étudiants à travailler au sein des offices nationaux de statistique a été posée en introduction au présent document. La réponse

[14] Dans son intervention lors de la conférence Morris Hansen, (2006), Michael Goodchild traite ce thème de façon très détaillée (Goodchild, 2007).

à cette question est généralement négative, mais la nécessité d'améliorer et d'intensifier la formation a en fait été posée dans diverses autres instances et des mesures positives sont d'ores et déjà en cours. Cette réflexion part de l'idée que personne ne remet en cause l'importance des offices nationaux de statistique et des organisations analogues et que les offices nationaux de statistique figurent parmi les plus gros pourvoyeurs d'emplois pour les statisticiens et les économistes.

40. Dans la présente section, les trois thématiques principales intéressant les travaux des offices nationaux de statistique seront abordées et la question de savoir si elles sont enseignées à l'université sera posée.

1. Échantillonnage des enquêtes

41. Il est évidemment inutile de s'attarder ici sur l'importance de l'échantillonnage des enquêtes dans le travail des offices nationaux de statistique. Il est impossible de concevoir une enquête, de nettoyer les données pour éliminer les valeurs incohérentes et les non-réponses, de produire des estimations appropriées et d'estimer les erreurs sans une solide connaissance de la théorie de l'échantillonnage. On présente ci-après le programme d'enseignement d'un cours d'une année intitulé « Conception et analyse des enquêtes par échantillonnage » proposé par l'Université de Harvard (trois heures par semaine) :

> *Les méthodes de conception et d'analyse des enquêtes par échantillonnage ; les caractéristiques de la boîte à outils de l'échantillonnage et leur utilisation dans le cadre de stratégies d'échantillonnage optimales ; les poids et méthodes d'estimation des variances, y compris les méthodes de rééchantillonnage ; un bref aperçu des aspects non statistiques de la méthodologie d'enquête tels que l'administration des enquêtes et les questionnaires.*

2. Ajustements saisonniers et estimation des tendances

42. Les séries chronologiques de données socioéconomiques sont utilisées pour étudier les tendances et détecter les évolutions (inversions de tendance) de l'activité socioéconomique. Or, ce processus d'apprentissage ne peut être mené à bien lorsque les séries chronologiques observées englobent non seulement le cycle tendanciel étudié, mais encore les variations saisonnières, les effets liés aux jours d'ouverture, les jours fériés flottants et les autres influences irrégulières. Pour pouvoir étudier la tendance, il convient alors d'estimer ces éléments supplémentaires, puis de les éliminer de la série observée. On obtient une première estimation de la tendance en soustrayant les effets saisonniers, une opération connue sous le nom d'ajustement saisonnier. Plusieurs procédures non paramétriques et basées sur des modèles ont été proposées dans la littérature pour l'ajustement saisonnier et l'estimation des tendances, et elles sont aujourd'hui couramment utilisées. En effet, beaucoup d'offices nationaux de statistique ont pour habitude de publier les séries de données corrigées des variations saisonnières ou les séries tendancielles en même temps que les séries originales observées. La différence entre les deux éléments tient au fait que les séries de données corrigées des variations saisonnières renferment également les termes irréguliers qui sont harmonisés pour estimer la tendance. Les séries tendancielles sont plus lisses, mais elles peuvent dissimuler des évolutions importantes, particulièrement vers la fin. Inversement, les séries de données corrigées des variations saisonnières peuvent faire apparaître des inversions de tendance erronées.

3. Comptabilité nationale

43. Un des principaux aspects du travail d'un office national de statistique concerne en la production de statistiques économiques de qualité. On peut notamment citer, entre autres exemples, la comptabilité nationale, la balance des paiements, les statistiques sur les finances publiques, les statistiques sur les prix, les statistiques sur le commerce international, les comptes satellites (énergie, protection sociale, éducation et autres), et la comptabilité économico-environnementale. Le système de comptabilité nationale (SCN) est un des principaux systèmes fondés sur des séries, et sa production est régie par des normes internationales strictes. Il permet aux utilisateurs et aux décideurs de comprendre de manière approfondie l'activité économique et son évolution. En Israël comme dans de nombreux autres pays, les analystes, les décideurs, les médias et le public attendent avec

une grande impatience chaque nouvelle publication d'estimations relatives aux comptes nationaux. On serait donc en droit de penser qu'un fondement aussi important de la statistique macroéconomique est enseigné dans chaque département d'économie de chaque université, mais est-ce bien le cas ?

B. Réponse à la question

44. Les universités enseignent-elles aux étudiants les connaissances de base dont ils ont besoin pour travailler dans ces trois domaines importants ? Pour répondre à cette question, nous avons parcouru les programmes de premier et second cycles en statistique et en économie des 25 meilleures universités du monde selon le Classement académique des universités mondiales, ou Classement de Shanghai (http://www.shanghairanking.com/) (voir l'annexe). Voici ce que nous avons appris :

45. Seules 11 des 25 universités les mieux classées proposent, à divers degrés, des cours d'initiation à l'échantillonnage d'enquête.

46. Si la plupart des universités dispensent un ou plusieurs cours consacrés à l'analyse et la prévision des séries chronologiques, aucune ne consacre une part significative du temps d'enseignement à l'ajustement saisonnier ou à l'estimation des tendances. En fait, seules trois universités mentionnent la saisonnalité dans le descriptif de leurs cours sur les séries chronologiques.

47. Les enseignements spécialisés consacrés à la comptabilité nationale sont encore plus rares. Dans le meilleur des cas, la comptabilité nationale n'apparaît qu'occasionnellement dans les cours de macroéconomie, principalement dans le contexte du produit intérieur brut (PIB). Il semble que seuls le Fonds monétaire international et les institutions qui en dépendent dispensent un enseignement détaillé sur la comptabilité nationale en plus de quelques programmes de *Master en statistique officielle* (voir plus loin).

48. À la lecture de ces conclusions, la question qui s'impose à l'esprit est la suivante : quel est en fait le rôle des universités publiques et privées ? Sont-elles censées former les étudiants et leur apprendre à travailler pour le compte d'un employeur, ou doivent-elles exclusivement se concentrer sur la recherche et former de nouvelles générations de chercheurs ? Ce débat, vieux de plusieurs siècles, dépasse bien évidemment le cadre du présent document. Il faudrait cependant s'intéresser de près aux points suivants :

a) Les trois thématiques mentionnées plus haut, ainsi que beaucoup d'autres thèmes sous-tendant le travail des offices nationaux de statistique, requièrent de solides connaissances en théorie statistique ou en sciences économiques. L'échantillonnage des enquêtes, l'ajustement saisonnier et l'estimation des tendances font appel à des notions complexes de statistique théorique dont les applications sont importantes. Un enseignement de ces thèmes n'irait donc pas à l'encontre de l'idée selon laquelle les universités devraient se concentrer sur la recherche. Certains des statisticiens mathématiciens les plus renommés au monde font des recherches sur les estimations par petites zones, autre domaine de travail très important des offices nationaux de statistique. Voici une liste non exhaustive d'exemples de thèmes intéressant les travaux des offices nationaux de statistique et basés sur la théorie statistique classique : méthodes d'échantillonnage des enquêtes, adaptation des modèles aux données d'enquêtes complexes, procédures de couplage, méthodes de contrôle de divulgation des statistiques, extraction de signaux des modèles ARIMA, estimation de l'erreur quadratique moyenne des estimations en données corrigées des variations saisonnières, utilisation de méthodes de rééchantillonnage pour réduire les écarts et estimer la variance ;

b) La pénurie de formations dans les domaines concernant les travaux des offices nationaux de statistique est peut-être due au manque de chercheurs spécialisés capables de dispenser cet enseignement. C'est pourquoi les étudiants ne sont pas exposés, pendant leurs études, au problème de l'échantillonnage des enquêtes et à d'autres problèmes importants que les offices nationaux de statistique rencontrent dans leur travail, et c'est aussi pourquoi ils ne prennent pas ces problèmes en considération dans leurs thèses de doctorat ou, par la suite, dans leurs projets de recherche universitaire ;

c) Depuis une dizaine d'années environ, les statistiques traversent de profonds bouleversements caractérisés par un glissement de ce qu'on appelle la « statistique classique » vers des méthodes nouvelles à forte intensité informatique fondées sur l'analyse des mégadonnées et autres notions comparables. L'échantillonnage des enquêtes et l'analyse des séries chronologiques ont aussi évolué, et il est évident que les enseignements de ces matières doivent être restructurés de façon à prendre en compte les évolutions modernes survenues dans ces différents domaines et, il faut l'espérer, à les rendre plus attractifs ;

d) Après tout, quelques universités dans différents pays mettent l'accent sur l'échantillonnage des enquêtes dans leurs enseignements et leurs activités de recherche. De plus, plusieurs universités proposent des programmes de *Master en statistique officielle*. Le programme conjoint en méthodologie des enquêtes proposé par les États-Unis et Westat et le programme de *Master en statistique officielle* de l'Université de Southampton (Royaume-Uni) sont bien connus à cet égard. Le premier est le fruit d'une collaboration entre l'Université du Maryland et l'Université du Michigan, et le second est financé par l'Office for National Statistics du Royaume-Uni. L'Institut national de la statistique et des études économiques (INSEE), en France, et l'Institut brésilien de géographie et de statistiques (IBGE) ont créé des écoles et instituts de statistique proposant des programmes spéciaux de licence et de *Master en statistique officielle*.

e) Enfin, l'Union européenne a récemment décidé de créer un programme européen de *Master en statistiques officielles* (EMOS), et elle a d'ores et déjà émis les appels d'offres auprès des offices nationaux de statistique et des universités d'Europe. Le *Master européen en statistiques officielles* (EMOS) est un réseau de programmes de master qui dispense un enseignement de second cycle en statistiques officielles à l'échelle européenne. EMOS est un programme conjoint qui regroupe des universités et des producteurs de données européens. Après deux appels à manifestation d'intérêt, plus de 20 programmes répartis dans 14 pays ont rejoint le réseau.

f) EMOS a été créé pour renforcer la coopération entre universités et producteurs de statistiques officielles et contribuer à former des professionnels capables de travailler avec les données officielles européennes de différents niveaux dans le système de production de statistiques du XXI^{e} siècle en rapide évolution. Le *Master EMOS* s'appuie sur l'enseignement des résultats qui permettent aux étudiants de se familiariser avec le système des statistiques officielles, les modèles de production, les méthodes statistiques et la diffusion des statistiques.

g) Les universités qui participent au programme de *Master EMOS* collaborent activement avec les offices nationaux de statistique pour réduire l'écart entre théorie et pratique. Les universités prennent par conséquent de plus en plus conscience de la nécessité de produire des statistiques officielles, et cette prise de conscience pourrait donc gagner d'autres universités.

VII. References[15]

[*Anglais seulement*]

AAPOR (2010). *Report on online survey panels*. http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.html?

AAPOR (2015). *Report on big data.* American Association for Public Opinion Research. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Abowd, J.M., and Vilhuber, L. (2008). *How protective are synthetic data?* In: PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin). Springer LNCS 5262, 239-246.

Cavallo, A., and Rigobon, R. (2010). *The Billion Prices Project@MIT*. (http://bpp.mit.edu).

Cavallo, A. (2012). *Online vs official price indexes: measuring Argentina's inflation.* Journal of Monetary Economics, 1-14.

Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). *A conditional empirical likelihood approach to combine sampling design and population level information.* Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

Couper, M.P. (2000). *Web surveys, a review of issues and approaches*. Public Opinion Quarterly 64, 464-494.

Couper, M.P. (2008). *Designing Effective Web Surveys.* Cambridge University Press.

Daas P.J.H., Puts M.J. Buelens B. and van den Hurk, P.A.M. (2013*). Big Data and official statistics. Proceedings of the NTTS*, Euro Stat, Brussels. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf

Dandekar, R.A., and Cox L.H. (2002). *Synthetic tabular data: an alternative to complementary cell suppression.* Manuscript, Energy Information Administration, U. S. Department of Energy.

De Leeuw, E. (2005*). To mix or not to mix? Data collection modes in Surveys*. Journal of Official Statistics, 21, 1-23.

Dillman, D.A., and Christian, L. (2005). *Survey mode as a source of instability in response across surveys.* Field Methods, 17, 30-52.

Dinur, I., and Nissim, K. (2003). *Revealing Information While Preserving Privacy.* PODS 2003, 202-210.

Dwork, C., McSherry, F. Nissim, K. and Smith, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC* (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Fay, R.E., and Herriot, R. (1979). *Estimates of income for small places: an application of James–Stein procedures to census data.* Journal of the American Statistical Association, 74, 269–77.

Feder, M., and Pfeffermann, D. (2015). *Statistical inference under non-ignorable sampling and nonresponse- an empirical likelihood approach.* Southampton Statistical Sciences Research Institute, http://eprints.soton.ac.uk/id/eprint/378245

Goodchild, M.F. (2007). *The Morris Hansen Lecture 2006: Statistical perspectives on social science.* Journal of Official Statistics, 23, 1–15.

Hartley, H.O., and Rao, J.N.K. (1968). *A new estimation theory for sample surveys*. Biometrika, 55, 547-557.

[15] References listed relate to those from the published article in the ***Journal of Survey Statistics and Methodology***, December 2015.

Lee, J., and Berger, J.O. (2001). *Semiparametric Bayesian Analysis of Selection Models*. Journal of the American Statistical Association, 96, 1397-1409.

Lee, S. (2006). *Propensity score adjustment as a weighting scheme for volunteer panel web surveys.* Journal of Official Statistics, 22, 329-349.

Lee, S., and Valliant, R. (2009). *Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment*. Sociological Methods & Research, 37. 319-343.

Little, R.J.A., and Liu, F. (2003). *Selective multiple imputation of keys for statistical disclosure control in microdata*. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.

National Research Council (2013). *Frontiers in Massive Data Analysis*. Washington D.C.: The National Academies Press. (http://www.nap.edu)

New Zealand (2012). *Using cellphone data to measure population movements*. (http://www.stats.govt.nz/tools_and_services/earthquake-info-portal/using-cellphone-data-report.aspx).

Nirel, R. and Glickman, H. (2009*). Sample surveys and censuses.* In: Handbook of Statistics 29A. Sample Surveys: Design, Methods and Application Eds., D. Pfeffermann and C.R. Rao. Amsterdam: North Holland, 539-565.

O'Keefe, C.M. and Good, N. (2008*). A remote analysis server – What Does Regression Output Look Like?* In PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 270-283.

O'Keefe, C.M. and Shlomo, N. (2012). *Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data*. Transactions on Data Privacy, 5, 403-432.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd edition). New York: Cambridge University Press.

Pfeffermann, D. (2013*). New important developments in small area estimation*. Statistical Science, 28, 40-68.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998). *Parametric distributions of complex survey data under informative probability sampling*. Statistica Sinica, 8, 1087-1114.

Pfeffermann, D. and Landsman, V. (2011*). Are private schools better than public schools?* Appraisal for Ireland by methods for observational studies. The Annals of Applied Statistics, 5, 1726–1751.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P.L. (2006). *Multilevel modeling under informative sampling.* Biometrika, 93, 943-959.

Pfeffermann, D. and Sikov A. (2011). *Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information.* Journal of Official Statistics, 27, 181-209.

Pfeffermann, D. and Sverchkov, M. (1999). *Parametric and semi-parametric estimation of regression models fitted to survey data.* Sankhya, 61, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003). *Fitting generalized linear models under informative probability sampling.* In: Analysis of Survey Data, eds. R. L. Chambers and C. J. Skinner, New York: Wiley, pp. 175-195.

Pfeffermann., D. and Sverchkov, M. (2007). *Small area estimation under informative probability sampling of areas and within the selected areas*. Journal of the American Statistical Association, 102, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ.MR1953089

Reiter, J.P. (2005). *Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical study.* Journal of the Royal Statistical Society, A, 168, 185-205.

Rivers, D. (2007). *Sampling for web surveys*. Joint Statistical Meeting, Proceedings of the Section on Survey Research Methods, Salt Lake City, UT, USA.

Rosenbaum, P.R. and Rubin, D.B. (1983). *The central role of the propensity score in observational studies for treatment effects.* Biometrika, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). *Reducing bias in observational studies using subclassification on the Propensity score.* Journal of the American Statistical Association, 79, 516-524.

Rotnitzky, A. and Robins, J. (1997). *Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response*. Statistics in Medicine, 16, 81-102.

Shlomo, N., Antal, L. and Elliot, M. (2015). *Measuring disclosure risk and data utility for flexible table generators*. Journal of Official Statistics, 31, 305–324.

Smith T.M.F. (1994). *Sample surveys 1975-1990; an age of reconciliation*? International Statistical Review, 62, 5-19.

Statistics and Science (2013). *A report of the London workshop on the future of the statistical sciences.* http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Sverchkov, M., and Pfeffermann, D. (2004). *Prediction of finite population totals based on the sample distribution.* Survey Methodology, 30, 79-92.

UN (2014). *Report of the global working group on big data for official statistics.* United Nations, E/CN.3/2015/4. http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf

Waksberg, J. and Goldfield, E. D. (1996). Morris Howard Hansen, 1920-1990. *A biographical memoir.* National Academy of Sciences, Washington D.C., U.S.A.

Vaccari, C. (2014). *Big Data in Official statistics. School of advanced studies*, University of Camerino, Italy. https://www.academia.edu/7571682/PhD).

Vannieuwenhuyze, J.T.A; Loosveldt, G and Molenberghs, G. (2014). *Evaluating mode effects in mixed-mode survey data using covariate adjustment models*. Journal of Official Statistics, 30, 1-21.