Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Science des données pour les systèmes de données dynamiques : Les implications pour la statistique officielle

Mary E. Thompson[1]

Résumé

Nombre des possibilités et des défis de la science des données moderne découlent d'éléments dynamiques, dont l'évolution des populations, la croissance du volume de données administratives et commerciales sur les particuliers et les établissements, les flux continus de données et la capacité de les analyser et de les résumer en temps réel, ainsi que la détérioration des données faute de ressources pour les tenir à jour. Le domaine de la statistique officielle, qui met l'accent sur la qualité des données et l'obtention de résultats défendables, se prête parfaitement à la mise en relief des questions statistiques et liées à la science des données dans divers contextes. L'exposé souligne l'importance des bases de sondage de population et de leur tenue à jour, la possibilité d'utiliser des méthodes à bases de sondage multiples et des couplages d'enregistrements, la façon dont l'utilisation de données à grande échelle non issues d'enquêtes comme information auxiliaire façonne les objets de l'inférence, la complexité des modèles pour les grands ensembles de données, l'importance des méthodes récursives et de la régularisation, et les avantages des outils évolués de visualisation des données en ce qui concerne la détection des changements.

Mots-clés : Données à grande échelle; combinaison de sources de données; méthodes récursives; réduction de la dimension; visualisation.

1. La nature dynamique de la statistique officielle

J'ai choisi pour sujet les « systèmes de données dynamiques » parce que le thème du symposium est « Croissance de l'information statistique : défis et avantages », et que la croissance est un concept dynamique. Même si nous définissons la statistique officielle très simplement comme consistant à « décrire l'état des choses » dans la société et l'économie, elle a toujours été dynamique, parce que l'« état des choses » évolue constamment, de manière prévisible et imprévisible.

J'aimerais considérer une définition assez générale de la statistique officielle, à savoir la production de sommaires des données recueillies par les organismes gouvernementaux ou au nom de ceux-ci aux fins de gouverner. Les données sont recueillies sous des auspices officiels; elles sont recueillies avec minutie, conformément à des plans et à des protocoles de collecte de données rigoureux; et elles sont censées être à la disposition (du moins sous forme de sommaires) des utilisateurs, assorties de mesures de la qualité. Aux termes de cette définition très générale, la statistique officielle pourrait être envisagée comme fournissant un portrait changeant de l'État et du monde, portrait qui englobe la population, l'économie, la santé, l'environnement et la société proprement dite. En fait, récemment, d'importants projets sociétaux et de justice sociale ont commencé à considérer les données comme un élément important de leurs travaux. La « révolution des données » pour les objectifs de développement durable (Nations Unies, 2014) et les travaux du Human Rights Data Analysis Group, p. ex., Price et Ball (2015), en sont des exemples. Au Canada, dans le rapport de la Commission de vérité et de réconciliation de 2015, l'appel à l'action nº 55 tourne autour de la fourniture de données.

Au sens de cette définition générale, la statistique officielle a été dynamique de nombreuses façons. Un élément important de son travail a toujours été le dépistage et la surveillance des changements dans les circonstances sociales et économiques de l'État. Un autre élément a été la construction et l'amélioration continue de bases de sondage pour les établissements, les régions et les ménages. Les plans d'échantillonnage ont été remaniés et rafraîchis régulièrement, et les estimations et les prévisions, mises à jour. Pendant des décennies, la statistique officielle a dû gérer de manière

[1] Mary E. Thompson, Université de Waterloo, Waterloo (Ontario) Canada, N2L 3L9

responsable l'accumulation de données, ainsi que les politiques de conservation et de diffusion des données. Et pourtant, en raison de grands changements technologiques et culturels, la statistique officielle doit maintenant faire face à une évolution plus rapide.

2. L'évolution de la conjoncture

Le Symposium 2016 a mis en vedette de nombreux aspects de l'évolution des opérations et de leurs principes directeurs. Par exemple, une discussion a porté sur l'usage croissant de données administratives dans les recensements. La conception des recensements, à une époque réalisés par vagues à intervalles de cinq ou de dix ans, est aujourd'hui de plus en plus une opération continue. L'article publié récemment par Blumerman (2015) au sujet des préparatifs du Recensement de 2020 aux États-Unis aborde certaines questions similaires.

Le rythme de la production de statistiques s'est accéléré, et pourrait s'accroître encore davantage. Nous avons été habitués depuis longtemps à ce que les chiffres économiques et financiers finaux sortent plusieurs mois après la période de référence qu'ils décrivent. En raison d'une plus grande automation de la collecte des données, nous pourrions voir certains de ces chiffres produits, comme l'ont souligné plusieurs conférenciers, presque en temps réel, ce qui facilitera la mise en place dans le domaine économique de « systèmes d'alerte rapide », et accroîtra la possibilité d'utiliser l'information pour modifier ou contrôler l'« état des choses ».

Naturellement, le traitement automatique présente des dangers, comme la « correction de cap » instantanée sur la base de fluctuations aléatoires, et l'induction de rétroactions, comme les réactions des marchés du travail en réponse à l'évolution des taux de change.

Les bases de sondage deviennent plus dynamiques en ce sens qu'elles font l'objet de mises à jour de plus en plus fréquentes; des exemples bien connus sont le Registre des entreprises de Statistique Canada, qui est mis à jour continuellement au moyen de données fiscales et d'autres sources, et le Delivery Sequence File des services postaux des États-Unis, qui est la base pour de nombreux plans de sondage fondés sur les adresses et qui est mis à jour hebdomadairement ou mensuellement. Dans le programme de la base de sondage pour les enquêtes auprès des ménages de Statistique Canada, qui est décrit dans un article du Symposium, les composantes de la base de sondage sont maintenant rafraîchies au moins trimestriellement en se servant de plusieurs sources.

Les bases de sondage tenues à jour par les organismes statistiques non seulement suivent l'évolution des populations, mais deviennent aussi plus riches, contiennent plus d'information auxiliaire, y compris de l'information sur les liens entre les unités. Elles deviennent plus appariables, ce qui se traduit par un vaste élargissement de leur utilité. Au moins à des fins administratives, il existe maintenant des bases de données qui peuvent être appariées en se servant des numéros d'identification d'étudiant, des numéros de carte santé, des numéros d'assurance sociale, d'images et même de l'ADN. Plusieurs explorations intéressantes des défis et des possibilités du couplage d'enregistrements ont été présentées durant le Symposium. Malgré l'usage prometteur et croissant d'identificateurs uniques, il existe encore certaines frontières, à savoir les applications où il est nécessaire de faire le meilleur usage possible d'identificateurs partiels traditionnels, tels que les noms, les adresses et les dates de naissance. Un article intéressant publié par Fu et coll. (2014) décrit l'élaboration d'une méthodologie pour le couplage automatique des personnes par la voie du couplage des ménages dans les données historiques de recensement.

Les processus de conception d'enquête deviennent délibérément dynamiques, faisant appel à des plans d'échantillonnage adaptatifs et à des plans de collecte des données dynamiques. Aujourd'hui, ces plans peuvent s'appuyer sur des paradonnées disponibles instantanément, par exemple pour déterminer les stratégies de suivi et les modes de collecte des données. Les nouveaux processus de conception présentent de nouvelles difficultés analytiques, par exemple l'intégration des effets du moment des interviews et des effets de mesure liés au mode de collecte dans les analyses.

Les nouvelles sources de données auxiliaires sont devenues abondantes, et la plupart des travaux de recherche récents, y compris certains décrits durant le Symposium, sont axés sur la façon d'utiliser ces nouvelles sources de manière économique. Les données sur les prix sont l'un des principaux exemples, d'importants progrès étant réalisés en ce qui concerne l'utilisation stratégique de données de lecteurs optiques et d'autres flux de données générés automatiquement.

En raison des nouvelles sources de données sur les prix, le concept d'indice de prix pourrait également évoluer. Pendant de nombreuses années, un indice des prix à la consommation (IPC) type a été construit à partir de deux sources de données d'enquête, à savoir les enquêtes auprès des établissements de détail (points de vente) et les enquêtes sur les dépenses des familles. Mais nous semblons être sur le point de pouvoir estimer le mouvement d'une quantité comme il suit :

$$\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{m_{jk}} p_{jki}\right)/J$$

où conceptuellement $j = 1, \ldots, J$ désigne les ménages, $k = 1, \ldots, K$ désigne les biens et services dans un « panier », et p_{jki} est le prix payé par le ménage *j* pour son *i*e achat de l'article *k* durant une année donnée. Cette quantité, la dépense globale par ménage durant l'année pour les biens et services du panier, n'est pas un indice de prix au sens classique, mais est étroitement reliée. Le « panier » pourrait être le panier de biens et services de l'IPC, ou une catégorie de produit, et la quantité est le prix moyen payé par le ménage durant l'année pour les biens et services dans le « panier ». Le numérateur peut être considéré comme une somme sur une population dont l'unité élémentaire est un achat. Au moins pour certains types d'articles, on disposera de bases de données sur les transactions d'achat.

Les méthodes d'utilisation des données auxiliaires sont réexaminées et adaptées, par exemple dans le cas de l'utilisation de flux de données « organiques » pour les « prévisions immédiates » des séries temporelles économiques (Castle et coll., 2015), ou dans celui de l'utilisation de bases de données sur le crédit et autres bases de données de ce type pour la correction des échantillons non probabilistes (Rivers, 2007).

3. La croissance des données et l'inférence

Étant donné l'évolution de la conjoncture qui vient d'être décrite, quelles sont les implications des données dynamiques et de la croissance des données en ce qui a trait à l'inférence et à la pratique de la statistique officielle? Ces implications dépendent à la fois des caractéristiques des données et de la nature de l'inférence.

Les nouveaux flux de données ont des caractéristiques variables. Certains ont une faible dimensionnalité, tandis que d'autres sont riches et de haute dimensionnalité. Certains sont très complets, la collecte des données étant serrée temporellement ou spatialement, et la fréquence de saisie, élevée. Dans de nombreux cas, les flux arrivent en « temps réel ». Certaines données sont d'une très grande exactitude, comme dans le cas des données de capteur, à condition que ce dernier fonctionne correctement. D'autres données peuvent être remplies d'erreurs, de lacunes ou de trous.

Qu'est-ce que l'inférence en statistique officielle? Il s'agit en grande partie d'une description, qui s'appuie principalement sur l'inférence fondée sur le plan de sondage, mais en utilisant des modèles pour améliorer l'efficacité ou tenir compte des erreurs non dues à l'échantillonnage. Le champ de la description englobe la surveillance en vue de déceler les anomalies. D'autres objectifs comprennent la prédiction ou la prévision, qui font appel à la modélisation pour relier le futur au passé et au présent. Enfin, la modélisation joue un rôle important dans la compréhension des connexions et des liens. Donc, en dépit de l'accent mis sur la description, l'inférence en statistique officielle est étroitement associée à l'inférence dans d'autres domaines de la statistique, et les modèles y occupent une place prédominante.

En ce qui concerne la caractérisation des défis statistiques, j'en cernerais trois, qui correspondent approximativement aux trois types d'inférences :

- Même si les objectifs descriptifs sont simples, les modèles bien ajustés aux ensembles de mégadonnées sont complexes.
- La prédiction et la prévision à partir de données qui s'accumulent requièrent une saisie en continu et des techniques d'inférence en continu.
- Des données de grande dimensionnalité et des modèles complexes demandent des techniques de réduction de la dimension aux fins de compréhension et de traitement.

3.1 Complexité des modèles

Dans un article récent intéressant, Cox (2015) soutient que [traduction] « les mégadonnées, comme il est convenu de les appeler, ont vraisemblablement une structure complexe, ce qui implique en particulier que les estimations de la précision obtenues en appliquant les méthodes statistiques classiques sont vraisemblablement trompeuses ». Il propose un modèle qui illustre l'effet de l'accumulation de sources de variabilité (p. ex., de l'échelon local à l'échelon national puis mondial) sur l'estimation des moyennes et des coefficients de régression; sous le modèle, la variance d'une moyenne ou d'un coefficient de régression diminue plus lentement que l'inverse de la « taille » de l'ensemble de données croissant.

Dans le cas de l'inférence descriptive habituelle, cette complexité ne pose pas problème. La population canadienne est grande et sa structure est complexe. Elle peut être décrite au moyen des données de recensement, et les caractéristiques de sa population active peuvent être estimées en utilisant l'inférence fondée sur le plan de sondage à partir de l'Enquête sur la population active menée auprès d'un échantillon probabiliste d'environ 56 000 ménages. Mais dans un contexte d'analyse, la complexité de la population est un élément très important. Si nous prenons un exemple extrême, on n'appliquerait pas un modèle de régression logistique simple à des données de recensement en s'attendant à ce que les erreurs-types quasi nulles et les résultats des tests des méthodes classiques reflètent ce que l'on comprend des dépendances dans les données.

3.2 Accumulation de données

Les plans d'échantillonnage dynamiques existent depuis longtemps, comme en témoignent les échantillons en continu (*rolling samples*) de Kish (1961; 1979; 1998); les plans de sondage pour un échantillon avec renouvellement tel que celui de Fellegi (1963); la technique de McLeod et Bellhouse (1983) pour tirer un échantillon aléatoire simple de taille n par « passage unique » à travers une population de taille inconnue N. L'utilisation de nombres aléatoires permanents (Ohlsson, 1995) permet de coordonner des échantillons successifs tirés d'une population de manière à répartir uniformément le fardeau de réponse. Ces concepts et techniques permettent essentiellement de gérer la croissance des données en laissant partir certaines de manière contrôlée, afin de maintenir un échantillon de taille approximativement constante.

3.3 Données de haute dimensionnalité et sélection du modèle

Les enquêtes des organismes de statistique officielle ont toujours été polyvalentes et habituellement conçues pour recueillir de multiples mesures sur les entreprises et les ménages. Des estimations doivent être produites pour des sous-populations définies par géographie, secteur ou groupes démographiques, et leur production est simple. Cependant, si l'on fait appel à la modélisation en vue d'améliorer l'efficacité ou comme moyen de traiter les données manquantes, le choix du modèle est un élément qui doit entrer en jeu. Il en est particulièrement ainsi des modèles pour la prévision, où la dimension de l'espace des variables explicatives ou prédictives ne cesse de s'accroître.

Les deux sections suivantes sont consacrées à une discussion plus détaillée de certaines méthodes applicables aux données dynamiques que la statistique officielle partage avec d'autres domaines. Les techniques récursives vont de pair avec les échantillons mobiles ou en continu, et les techniques de réduction de la dimension sont un outil important pour la sélection des modèles.

4. Inférences récursives

Les techniques d'inférence récursive, et plus généralement ce que nous pourrions appeler les méthodes d'estimation mobile ou en continu (*rolling*), dans lesquelles des algorithmes d'estimation ou de prévision des décalages temporels, sont bien établis en statistique officielle. Les méthodes ARIMA pour séries temporelles et leurs extensions en sont un exemple marquant. La production de modèles et de prévisions pour des séries temporelles de grande dimensionnalité ou réparties spatialement, idéalement en satisfaisant certaines exigences de cohérence, comme des contraintes d'agrégation, devient de plus en plus importante (Quenneville et Fortier, 2012). Hyndman et coll. (2016) proposent des algorithmes à matrices creuses pour rapprocher les prévisions pour de grands nombres de séries temporelles groupées.

D'autres méthodes sont reliées à des modèles espace-état, au filtre de Kalman et à ses généralisations, ou appliquées à des estimations de moyenne au cours du temps par des auteurs tels que Tam (1987) et Pfeffermann (1991); voir aussi Tam (2015). Voici un exemple simple de modèle espace-état pour une série temporelle y_t (un cas particulier du modèle à filtre de Kalman)

$$y_t = x_t\beta_t + u_t$$

où t désigne la période, l'état β_t est un processus AR(1), et les u_t sont indépendants et de loi $N(0, \sigma^2)$. Plus généralement, l'état β_t est un processus markovien, et la distribution de y_t est déterminée par l'état, ce qui rend possible l'estimation de l'état à chaque période *t* en se basant sur l'observation de la série temporelle jusqu'à ce point dans le temps. L'état de la série temporelle, qui pourrait par exemple être une moyenne de population mobile, est estimé afin d'éliminer le bruit de la série temporelle. Le but est de comprendre la variation de l'état ou du « signal » au cours du temps, de manière à pouvoir le résumer de façon convaincante et de bien le projeter.

Les modèles espace-état sont des modèles, mais il existe un analogue fondé sur le plan de sondage dans une application d'estimation composite, utilisée dans l'Enquête sur la population active du Canada, pour accroître l'efficacité et lisser la série temporelle composante (Singh et coll., 2001; Fuller et Rao, 2001). Cette application d'estimation composite par régression à une enquête avec plan de sondage à panel rotatif utilise les données du mois précédent pour les répondants qui persévèrent afin d'améliorer l'estimation du niveau et de la variation durant le mois courant. L'estimation composite ainsi que le filtre de Kalman peuvent être dérivés d'une combinaison efficace de fonctions dont l'espérance est nulle et en ce sens sont analogues.

L'approche espace-état est particulièrement utile quand une tendance spatiale ou temporelle correspond à un modèle complexe, basé sur la théorie physique ou biologique. On en trouve un exemple dans un article de Shaman et coll. (2014), dans lequel l'épidémie d'Ébola en Afrique occidentale a été modélisée au moyen d'un modèle SEIRX (susceptible-exposé-infectieux-rétabli-décédé) qui fournit la forme de l'évolution de l'état, qui est le nombre reproductif quotidien. Un filtre de Kalman généralisé utilise les observations hebdomadaires pour mettre à jour l'estimation de l'état, de manière à pouvoir projeter la situation de l'épidémie six semaines dans le futur. L'objectif est d'essayer de générer des prévisions réalistes et rapides de ce phénomène hautement non stationnaire.

D'autres exemples de l'utilisation des modèles espace-état comprennent la prévision saisonnière du rendement des cultures (Newlands et coll., 2014) et l'estimation de la maturité des stocks de poisson (Xu et coll., 2015), dans l'un et l'autre cas à des fins importantes pour les décideurs gouvernementaux.

5. Réduction de la dimension et régularisation

Étant donné la croissance des données disponibles, nous pouvons nous attendre à ce que la modélisation joue un rôle de plus en plus important dans la production de sommaires statistiques. Parallèlement, les données à modéliser deviendront plus complexes. Elles pourraient avoir une plus grande dimensionnalité, au sens où les mesures ont une grande dimensionnalité ou au sens où les sources de variation sont plus nombreuses. Donc, la réduction de la dimension deviendra de plus en plus importante en statistique officielle.

Un cas d'espèce est celui des données fonctionnelles, où un point de données est une fonction temporelle et, au sens conceptuel, peut donc avoir une dimension infinie. Un exemple intéressant est l'estimation de la courbe de consommation moyenne d'électricité sur l'ensemble des clients dans une région, où des millions de consommateurs, chacun équipé d'un compteur d'électricité, envoient des lectures de consommation faites à des échelles de temps de niveau très fin (Lardin-Puech et coll., 2014). Les moyens classiques de réduire efficacement la dimension comprennent l'analyse de Fourier ou d'autres façons d'approximer la fonction par une combinaison linéaire de fonctions de base. Une autre forme de réduction de la dimension serait l'interpolation linéaire de la fonction entre les valeurs à des points dans le temps échantillonnés.

La réduction de la dimension en vue de choisir un modèle parcimonieux pourrait devenir de plus en plus importante à mesure que nous essayons d'exploiter des données plus riches pour pallier les données manquantes. Phipps et Toth

(2012) ont appliqué des arbres de régression pour trouver un modèle parcimonieux de la propension à répondre dans une enquête auprès des établissements dotée d'une base de sondage relativement riche. Un problème intéressant se pose, à savoir la façon d'utiliser la même sorte de données auxiliaires pour l'imputation.

La régularisation est une généralisation de la réduction de la dimension des paramètres en vertu de laquelle un objet d'inférence dont la représentation est instable, tel qu'une densité, est approximé par un objet ayant une représentation plus stable (Bickel et Li, 2006). Il s'agit d'une technique permettant d'éviter le surajustement. Elle possède des applications dans le domaine des séries temporelles où l'objet d'inférence serait la structure d'autocorrélation ou une densité spectrale. Elle permet à Bickel et Gel (2011) d'approximer une série temporelle éventuellement non linéaire par une « longue série autorégressive », et à Burr et coll. (2015) à Santé Canada de modéliser les associations entre des séries temporelles de statistiques sur l'environnement et la mortalité. Bornn et Zidek (2012) utilisent la régularisation bayésienne dans la modélisation spatiale des rendements des cultures au cours du temps dans les Prairies canadiennes, aux fins de prédiction par des variables liées au climat et liées au sol.

6. Visualisation des données

Bien que la représentation visuelle de l'information en statistique puisse souvent être trompeuse, la visualisation des données pourrait occuper une place croissante en statistique officielle, pour l'exploration des données et pour l'amélioration de la communication.

Dans d'autres domaines, la visualisation des données possède d'importantes applications de surveillance et de contrôle. Par exemple, des ingénieurs travaillant dans les secteurs de la fabrication et des produits chimiques ont mis au point des techniques avancées faisant appel non seulement à la visualisation à l'« œil nu », mais aussi à l'analyse automatique des images (Duchesne et coll., 2012). Des données réelles ou synthétiques sur la congestion de la circulation peuvent être superposées à des cartes terrestres Google (Kwoczek et coll., 2014). Il est possible d'imaginer un grand nombre d'applications de ce genre où la visualisation peut aider à décider comment, où et quand intervenir.

À une époque où l'on s'attendait fortement à une épidémie de grippe, le projet SIMID a produit un tableau de bord prototype pour surveiller la progression d'une épidémie de grippe dans la région de Peel en Ontario (Ramírez Ramírez et coll., 2012). Nous avons construit un réseau de contacts hypothétiques (familles, écoles et lieux de travail) et exécuté des microsimulations d'un modèle stochastique d'épidémie sur le réseau, montrant la progression de chacun sur une carte de la région pour diverses valeurs des paramètres de l'épidémie, tels que la période de latence et les paramètres de contrôle, tels que le taux de vaccination.

La cartographie peut également faciliter la représentation des données d'enquête et de population, y compris les strates et les unités du plan de sondage, en vue de l'exploration des données. Par exemple, quand les résultats d'intérêt sont reliés à la géographie, il peut être utile de superposer les strates et les unités primaires d'échantillonnage d'une enquête sur une carte de la densité de population provenant du recensement.

L'estimation sur petits domaines fait maintenant usage de « sources de mégadonnées » (Marchetti et coll., 2015), et la visualisation des moyennes de petit domaine sur des cartes est de plus en plus souvent possible, sous réserve des mesures de prudence habituelles. Le US Census Bureau possède un site Web consacré aux estimations de l'assurance-maladie par petit domaine pour 2005 à 2013, sur lequel sont animés les résultats de l'estimation sur petits domaines au niveau du comté en utilisant l'interpolation de splines à l'intérieur des États à partir des centres géométriques des comtés. Sangalli et coll. (2013) se servent aussi d'une technique d'interpolation de splines pour représenter les données de recensement sur la densité de population à Montréal.

7. Conclusion

Le présent article passe en revue les méthodes reliées à la croissance des données et à certains travaux présentés dans le cadre du Symposium. Il est clair que nous progressons rapidement pour ce qui est d'exploiter le potentiel des nouvelles sources de données. Les méthodes de collecte et de gestion des données ont changé radicalement, suscitant

de nouvelles possibilités et de nouveaux défis analytiques. L'évolution des méthodes d'analyse mène à de plus nombreux points de contact entre les pratiques naissantes et d'autres branches de la statistique et de l'informatique.

Bibliographie

Bickel, P. J. et Gel, Y. R. (2011), "Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series", *Journal of the Royal Statistical Society Series B*, 73, pp. 711-728.

Bickel, P. G. and Li, B. (2006), "Regularization in statistics", *Test*, 15, pp. 271-344.

Blumerman, L. (2015), "Planning for the 2020 census: A new design for the 21st century", *Amstat News*, Issue 462, pp. 12-13.

Bornn, L. and Zidek, J. V. (2012), "Efficient stabilization of crop yield prediction in the Canadian Prairies", *Agricultural and Forest Meteorology*, 152, pp. 223-232.

Burr, W., Takahara, G. and Shin, H. H. (2015), "Bias correction in estimation of public health risk attributable to short-term air pollution exposure", *Environmetrics*, 26, pp. 298-311.

Castle, J. L., Hendry, D. F. and Kitov, O. (2015), "Forecasting and nowcasting macroeconomic variables: a methodological overview", *Handbook on Rapid Estimates*, Eurostat.

Cox, D. R. (2015), "Big data and precision", *Biometrika*, 102, pp. 712-716.

Duchesne, C., Liu, J. J. and MacGregor, J. F. (2012), "Multivariate image analysis in the process industries: a review", *Chemometrics and Intelligent Laboratory Systems*, 117, pp. 116-128.

Fellegi, I. (1963), "Sampling with varying probabilities without replacement: rotating and non-rotating samples", *Journal of the American Statistical Association*, 58, pp. 183-201.

Fu, Z., Boot, H. M., Christen, P. and Zhou, J. (2014), "Automatic record linkage of individuals and households in historic census data", *International Journal of Humanities and Arts Computing*, 8, pp. 204-225.

Fuller, W. A. and Rao, J. N. K. (2001), "A regression composite estimator with application to the Canadian Labour Force Survey", *Survey Methodology*, 27, pp. 45-51.

Hyndman, R. J., Lee, A. J. and Wang, E. (2016), "Fast computation of reconciled forecasts for hierarchical and grouped time series", *Computational Statistics and Data Analysis*, 97, pp. 16-32.

Kish, L., Lovejoy, W. and Rackow, P. (1961), "A multi-stage probability sample for traffic surveys", *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 227-230.

Kish, L. (1979), "Samples and censuses", *International Statistical Review*, 47, pp. 99-109.

Kish, L. (1998), "Space/time variations and rolling samples", *Journal of Official Statistics*, 14, pp. 31-46.

Lardin-Puech, P., Cardot, H. and Goga, C. (2014), "Analysing large sets of functional data from a survey sampling point of view," *Journal de la Société Française de la Statistique*, 155, pp. 70-94.

Kwoczek, S., Di Martino, S., Nejdl, W. (2014), "Predicting and visualizing traffic congestion in the presences of planned special events", *Journal of Visual Languages and Computing*, 25, pp. 973-980.

McLeod, A. I. and Bellhouse, D. R. (1983), "A convenient algorithm for drawing a random sample", *Applied Statistics*, 32, pp. 182-184.

Newlands, N. K., Zamar, D. S., Kouadio, L. A., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S. and Hill, H. S. J. (2014), "An integrated, probabilistic model for improved seasonal forecasting of crop yield under environmental uncertainty", *Frontiers in Environmental Science*, 2, pp. 1-21.

Ohlsson, E. (1992), *SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden – A Methodological Description*, Stockholm, Sweden: Statistics Sweden.

Pfeffermann, D. (1991), "Estimation and seasonal adjustment of population means using data from repeated surveys", *Journal of Business and Economic Statistics*, 9, pp. 163-175.

Phipps, P. and Toth, D. (2012), "Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data", *Annals of Applied Statistics*, 6, pp. 772-794.

Price, M. and Ball, P. (2015), "Selection bias and the statistical patterns of mortality in conflict", *Statistical Journal of the IAOS*, 31, 263-272.

Quenneville, B. and Fortier, S. (2012), "Restoring accounting constraints in time series: methods and software for a statistical agency", In: W. R. Bell et al. (eds) *Economic Time Series: Modeling and Seasonality*, Boca Raton, Florida: CRC Press, pp. 231-253.

Ramírez-Ramírez, L. L., Gel, Y. R., Thompson, M., de Villa, E. and McPherson, M. (2012), "SIMID: SIMulation of Infectious Diseases using Random Networks", *Computer Methods and Programs in Biomedicine*, 110(3), pp. 455-470.

Rivers, D. (2007), "Sampling for web surveys". *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Sangalli, L. M., Ramsay, J. and Ramsay, T. (2013), "Spatial spline regression models", *Journal of the Royal Statistical Society Series B*, 75, pp. 681-703.

Shaman, J., Yang, W. and Kandula, S. (2014), "Inference and forecast of the current West African ebola outbreak in Guinea, Sierra Leone and Liberia", *PLOS Currents Outbreaks*, doi: 10.1371/currents.outbreaks.3408774290b1a0f2dd7cae877c8b8ff6.

Singh, A. C., Kennedy, B. and Wu, S. (2001), "Regression composite estimation for the Canadian Labour Force Survey with a Rotating Panel Design", *Survey Methodology*, 27, pp. 33-44.

Tam, S. M. (1987), "Analysis of repeated surveys using a dynamic linear model", *International Statistical Review*, 55, pp. 67-73.

Tam, S. M. (2015), "A statistical framework for analysing big data", *The Survey Statistician*, July 2015, pp. 36-51.

United Nations (2014), *A World that Counts: Mobilising the Data Revolution for Sustainable Development.* Report of the UN Secretary-General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development.

Wang, Q. and Rao, J. N. K. (2002), "Empirical likelihood-based inference under imputation for missing response data", *The Annals of Statistics*, 30, pp. 896-924.

Xu, X., Canton, E., Mills Flemming, J. and Field, C. (2015), "Robust state space models for estimating fish stock maturities", *Canadian Journal of Statistics*, 43, pp. 133-150.