Quelques réflexions sur la statistique officielle et son avenir

Yves Tillé [1]

[1] *Université de Neuchâtel, Bellevaux 51, 2000 Neuchâtel, Suisse, yves.tille@unine.ch*

Résumé. Dans cet article, nous partageons quelques réflexions sur l'état de la science statistique et son évolution dans les systèmes de production de la statistique publique. Nous tentons d'abord de faire une synthèse de l'évolution de la pensée statistique. Nous examinons ensuite l'évolution des pratiques de la statistique publique, qui a dû faire face très tôt à une diversification des sources : d'abord avec l'utilisation des recensements, puis des enquêtes par sondage et enfin des fichiers administratifs. A chaque étape, une profonde révision des méthodes a été nécessaire. Nous montrons que depuis le milieu du 20ème siècle, l'un des défis majeurs de la statistique est de produire des estimations à partir de sources variées. Pour ce faire, un grand nombre de méthodes ont été proposées, qui reposent sur des bases très différentes. Le terme *big data* englobe un ensemble de sources et de nouvelles méthodes statistiques. Nous examinons d'abord le potentiel de valorisation des *big data* dans la statistique publique. Certaines applications comme l'analyse d'images pour la prédiction agricole sont très anciennes et seront développées. Cependant, nous faisons part de notre scepticisme à l'égard des méthodes de dépouillement du web. Nous examinons ensuite l'utilisation de nouvelles méthodes d'apprentissage profond. Ces méthodes sont prometteuses mais soulèvent de nouvelles questions épistémologiques. Avec l'accès à de plus en plus de sources, le grand défi restera la valorisation et l'harmonisation de ces sources.

Mots-clés. apprentissage statistique, déduction, échantillonnage, fondements, induction, Lasso, registre, valeur p

Abstract. In this article, we share some reflections on the state of statistical science and its evolution in the production systems of official statistics. We first try to make a synthesis of the evolution of statistical thinking. We then examine the evolution of practices in official statistics, which had to face very early on a diversification of sources: first with the use of censuses, then sample surveys and finally administrative files. At each stage, a profound revision of methods was necessary. We show that since the middle of the 20th century, one of the major challenges of statistics has been to produce estimates from a variety of sources. To do this, a large number of methods have been proposed which are based on very different foundations. The term "big data" encompasses a set of sources and new statistical methods. We first examine the potential of valorization of big data in official statistics. Some applications such as image analysis for agricultural prediction are very old and will be further developed. However, we report our skepticism towards web-scrapping methods. Then we examine the use of new deep learning methods. These methods are promising but raise new epistemological questions. With access to more and more sources, the great challenge will remain the valorization and harmonization of these sources.

Keywords. deduction, foundations, induction, Lasso, p-value, registers, sampling, statistical learning

1 Introduction

La statistique officielle est un domaine un peu particulier de la statistique. Les méthodes qui y sont utilisées ont été développées pour répondre à des problèmes particuliers. En statistique officielle, la préoccupation principale n'est pas la prise de décision mais la qualité des estimations proposées. Dans cet article, nous essayons d'entrevoir l'avenir de la méthodologie statistique. Actuellement, la statistique est confrontée à de nombreuses questions épistémologiques, dont la plus emblématique est la crise de la valeur p. Nous décrivons les spécificités de la statistique officielle en retraçant l'histoire des sources et l'histoire des controverses autour de la méthodologie. Puis, nous analysons l'impact des nouvelles sources de données et des nouvelles méthodes statistiques. Une évolution de la méthodologie sera nécessaire et devra être soutenue par une recherche fondamentale de qualité.

2 La science statistique

La statistique est une science qui vise à étudier une réalité par le traitement, l'analyse, la modélisation et l'interprétation des données. Traditionnellement, il existe plusieurs approches statistiques.

1. *Statistique descriptive ou exploratoire*, qui consiste à présenter les données de manière plus condensée à l'aide de tableaux et de graphiques. L'objectif consiste à réduire la complexité des données au moyen de méthodes de réduction de dimensionnalité ou classification automatique. La statistique descriptive est avant tout liée à une démarche interprétative et exploratoire. John Wilder Tukey était un fervent partisan de la statistique descriptive. Son livre *Exploratory Data Analysis* (Tukey, 1977) reste une référence de l'approche exploratoire. De même, en France, l'école d'analyse des données initiée par Jean-Paul Benzécri a promu la statistique descriptive et exploratoire (Benzécri, 1973a,b; Bastin *et al.*, 1980).

2. *Statistique analytique ou inférentielle* qui vise à déduire des propriétés sur une population à partir de données. Cette population peut être fictive et être, par exemple, une distribution de probabilité ou un modèle. La statistique inférentielle repose entièrement sur des calculs de probabilité. Elle comprend la théorie des tests d'hypothèses statistiques et la théorie de la décision. Les principaux fondateurs de cette théorie sont Karl Pearson, William Sealy Gosset, Ronald Fisher et Jerzy Neyman.

3. *Modélisation* consiste à décrire la réalité par un modèle général décrit par une ou plusieurs équations. Un modèle est nécessairement une approximation plus simple que la réalité, comme l'écrit Box et Draper (2007) : "N'oubliez pas que tous les modèles sont faux ; la question pratique est de savoir jusqu'à quel point ils doivent être faux pour ne pas être utiles"[1]. Les modèles peuvent être utilisés soit pour décrire les relations entre les variables, soit pour faire des prédictions.

[1]Traduit de l'anglais "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful"

3 Interprétation des approches statistiques

On distingue généralement deux approches scientifiques : Le *approche inductive* est basé sur des observations et des données et vise à construire une théorie générale. La *approche déductive* (ou hypothético-déductive) est basée sur une théorie. Elle vise à déduire des résultats particuliers d'une théorie générale afin de vérifier si ces résultats particuliers peuvent être confirmés par des observations.

La statistique exploratoire peut être associée à une approche inductive. Dans cette approche, on part des données pour essayer de trouver empiriquement une explication globale. La statistique inférentielle, et plus particulièrement la théorie des tests d'hypothèses statistiques, peut être considérée comme une approche hypothético-déductive. On formule une hypothèse, puis on se prononce sur cette hypothèse en la confrontant aux données. Les deux approches sont qualifiées de complémentaires. L'analyse exploratoire permet de formuler des hypothèses. Ensuite, la statistique inférentielle peut éventuellement les confirmer. Un retour à l'analyse exploratoire permet enfin de formuler de nouvelles hypothèses et ainsi de suite. C'est ce que l'on appelle le cycle induction-déduction qui serait le moteur de la production de connaissances. Souvent, ce cycle a été complètement perverti, par exemple en utilisant les mêmes données pour formuler et tester des hypothèses, ce qui rend l'application de la théorie des tests complètement caduque. Cette approche est ironiquement appelée HARKing par Norbert L. Kerr (1998) dont l'acronyme vient de *Hypothesizing After the Results are Known*.

Cependant, cette vision de l'opposition induction-déduction est loin d'être unanimement partagée. Le rôle de l'induction a fait l'objet d'une controverse entre Jerzy Neyman et Ronald Fisher. Fisher (1935) a promu le raisonnement inductif. Jerzy Neyman (1957) avait une approche plus décisionnelle dans la théorie des tests d'hypothèse mais il était ensuite réticent à l'étendre au raisonnement inductif. La statistique inférentielle peut être considérée comme un raisonnement inductif car elle vise à extrapoler les résultats d'un échantillon à une population ou à un modèle. Il s'agit donc d'un processus de généralisation qui peut être considéré comme fondamentalement inductif (voir entre autres Lehmann, 1993; Capel *et al.*, 1996).

Costantini et Galavotti (1986) affirment que les méthodes d'estimation, telles que la méthode du maximum de vraisemblance, correspondent à une approche inductive et que la théorie des tests d'hypothèses correspond généralement à une approche déductive. D'autres interprétations existent. Par exemple, Adrew Gelman (2011) discute d'une philosophie de la statistique qui considère la statistique fréquentiste comme déductive et la statistique bayésienne comme inductive.

Face à ces différentes interprétations, on ne peut que souscrire à l'affirmation de Capel *et al.* (1996) qui écrit : "En tout état de cause, une compréhension réelle du rôle du raisonnement inductif dans les sciences humaines qui soit commune au statisticien, au chercheur et au praticien n'existe manifestement pas, et en particulier, comme nous l'avons vu, en ce qui concerne l'utilisation des tests d'hypothèse" : ‘

En outre, la distinction entre induction et déduction est parfois considérée comme dépassée. Karl Popper (2005) rejette la description du processus scientifique comme un cycle d'induction-

déduction. Il rejette carrément l'intérêt de l'approche inductive en science : "Pourtant, même en supposant que ce soit le cas – car après tout, "toute la science" pourrait se tromper – je continuerais à soutenir qu'un principe d'induction est superflue, et qu'il doit conduire à des incohérences logiques" [2]. Karl Popper soutient que la méthode scientifique consiste à formuler des propositions scientifiques qui doivent pouvoir être falsifiées par une expérience. En effet, aucune théorie ne peut être prouvée par une expérience. Le fait qu'un modèle soit compatible avec des données ne prouve jamais que le modèle est vrai. En effet, un autre modèle pourrait également être compatible avec le même ensemble de données.

Le cycle induction-déduction perd aussi complètement son sens dans certaines applications statistiques. En statistique officielle, cette distinction peut difficilement être appliquée. En effet, l'objectif consiste souvent à estimer certaines caractéristiques d'une population à partir de sources éparses (enquêtes, fichiers administratifs, recensements). L'objectif n'est donc pas d'établir une théorie scientifique mais simplement de pallier l'impossibilité d'obtenir une mesure complète et correcte sur toutes les unités de la population d'intérêt.

Ce cycle induction-déduction perd également son sens avec l'avènement de nouvelles méthodes dites d'apprentissage statistique (machine à vecteur de support, réseau de neurones, plus proche voisin, méthodes pénalisées, forêts aléatoires) qui permettent de prédire sans réellement modéliser. On pourrait considérer ces méthodes comme typiquement inductives car elles n'impliquent aucune formalisation a priori de la réalité (Harman et Kulkarni, 2012). Le terme *apprentissage statistique* est cependant un peu trompeur car ces méthodes ne conduisent pas vraiment à une théorisation qui serait une modélisation automatique des données. Le cycle dit d'induction/déduction est en quelque sorte contourné. Par exemple, si nous ne connaissons pas le revenu d'un individu dans une base de données, nous pouvons lui attribuer le revenu de l'unité statistique qui lui ressemble le plus. Il s'agit d'une prévision par la méthode du plus proche voisin. Peut-on pour autant dire que cette approche consiste à modéliser le revenu ? Tout au plus, on peut discuter de ce que signifie "ressembler" mais on est souvent obligé d'utiliser uniquement les variables disponibles dans les fichiers pour définir une distance entre les unités. Or, cette approche peut être très efficace. On ne voit pas bien comment on pourrait dire que cette approche est inductive ou déductive. Pour le statisticien, la principale question concernant ce type de méthode est d'évaluer et d'estimer sa précision. Les méthodes de réseaux de neurones ne permettent pas non plus de comprendre finement la relation entre les variables. Cependant, les méthodes d'apprentissage par arbre de décision permettent une post-interprétation. Avec les forêts aléatoires, il est également possible d'avoir une idée de l'importance des variables dans les prédictions.

Les nouvelles méthodes d'*apprentissage statistique* bousculent donc un principe de base de la démarche scientifique qui voudrait que toute connaissance puisse être transmise par le discours. Si les prévisions par *réseaux de neurones, plus proches voisins* ou *forêts aléatoires* peuvent très bien fonctionner pour faire une prédiction, elles ne permettent pas d'établir une théorie générale ou un principe qui nous permettrait de comprendre les relations entre les variables.

[2]Traduit de l'anglais "Yet even supposing this were the case – for after all, 'the whole of science' might err – I should still contend that a principle of induction is superfluous, and that it must lead to logical inconsistencies."

4 Crise de la valeur p

Dans les tests d'hypothèses statistiques, la valeur p est la probabilité que sous une hypothèse (conventionnellement appelée hypothèse nulle), nous obtenions la même valeur ou une valeur encore plus extrême que celle obtenue avec les données observées. Si la valeur p est faible, alors nous rejetons cette hypothèse. L'erreur de première espèce est définie comme la probabilité de rejeter l'hypothèse nulle étant donné qu'elle est vraie. Si nous effectuons le test avec une erreur de première espèce de 5% par exemple, nous rejetons l'hypothèse si la valeur p est inférieure à 5%.

La valeur p est devenue un argument décisif dans de nombreuses sciences : sciences humaines, économie, finance ou biologie. Cependant, dans les revues scientifiques de statistique, on peut être surpris du très faible nombre de valeurs p utilisées dans les articles publiés. La valeur p est souvent interprétée à tort comme la probabilité que l'hypothèse nulle soit vraie, ce qui n'est évidemment pas le cas.

Dans de nombreuses publications, plusieurs valeurs p (parfois des dizaines) apparaissent sans aucune réflexion sur la probabilité d'avoir au moins une valeur p inférieure à 5% si toutes les hypothèses nulles étaient vraies. Très souvent, les chercheurs effectuent un grand nombre de tests d'hypothèses et ne publient que les résultats pour les valeurs p inférieures à 5%. Cette approche est similaire à celle de HARKing. Une autre erreur méthodologique consiste à identifier un modèle en choisissant les variables ayant des valeurs p inférieures à 5% parmi un très grand nombre de variables. Cette procédure conduit inévitablement à une sur-spécification du modèle. Cependant, de nouvelles solutions ont été développées pour sélectionner les variables en utilisant par exemple la méthode Lasso (Tibshirani, 1996, 2011).

Ces utilisations des valeurs p ont été dénoncées dans un article au titre provocateur : *Why Most Published Research Findings Are False* (Pourquoi la plupart des résultats de recherche publiés sont faux) (Ioannidis, 2005). Le sujet est devenu controversé à tel point qu'on peut dire qu'il y a une crise de la statistique (Gelman et Loken, 2014; Fraser et Reid, 2016) et que l'*American Statistical Association* s'est sentie obligée de publier une déclaration sur la valeur p (Wasserstein et Lazar, 2016).

5 Qu'en est-il de la statistique officielle ?

La charte de la statistique officielle suisse est disponible sur le site du Conseil d'éthique de la statistique officielle (www.iecitepchartesuisse). Dans cette charte, la mission de la statistique publique est définie : "La statistique publique a pour mission de répondre aux besoins d'informations statistiques d'intérêt général de la société ainsi qu'à ceux relatifs à la conduite des politiques publiques.".

La mission ne consiste donc pas à interpréter, modéliser, établir des connaissances ou décider. Le cycle induction-déduction de la production de connaissances n'est donc pas un cadre approprié dans la statistique publique. En effet, en statistique publique, la démarche n'est ni exploratoire ni décisionnelle. Les méthodes statistiques développées sont spécifiques,

car il ne s'agit pas de faire de la recherche scientifique pour établir de nouvelles connaissances, mais de répondre aux besoins d'information en fournissant des statistiques fiables, durables et de qualité.

Il existe cependant un cycle entre la société civile et politique et la statistique publique afin de déterminer les besoins. La statistique publique doit assurer la continuité mais aussi renouveler sa production statistique en s'ouvrant à de nouveaux thèmes comme les inégalités de genre ou les préoccupations environnementales. Les méthodes statistiques utilisées dans la statistique publique ne sont donc ni exploratoires ni décisionnelles. Les statisticiens officiels ont mis l'accent sur la qualité. Un document exemplaire est le *Cadre d'assurance de la qualité de Statistique Canada* dont les principes de base sont la pertinence, l'exactitude, l'actualité, l'accessibilité, la cohérence, l'interprétabilité (Statistics Canada, 2017).

6 Les sources de données dans la statistique officielle et leur intégration

La statistique officielle a donc développé des méthodes un peu particulières pour atteindre leurs objectifs. L'avantage est que le statistique officielle est à l'abri de la crise de la valeur p car son utilisation est relativement limitée. L'histoire des méthodes de la statistique officielle peut être résumée en quelques ères (voir entre autres Hansen et Madow, 1974; Kruskal et Mosteller, 1980; Hansen, 1987; Bellhouse, 1988; Bethlehem, 2009; Tillé, 2020). Tout d'abord, il y a l'ère des recensements qui couvre tout le XIXe siècle. Durant cette période, seule la compilation exhaustive de données était considérée comme scientifique. Cette doctrine est clairement énoncée par le statisticien Adolphe Quételet (1846). La rupture est initiée par le directeur de l'Institut norvégien de statistique : Anders Nicolai Kiær (1896, 1899, 1903, 1905) qui propose d'utiliser des données partielles et donc un échantillon. Après une longue controverse, l'idée d'utiliser des échantillons a finalement été acceptée par l'Institut international de statistique (Jensen, 1926). Ken Brewer (2013) interprète ce débat comme la première controverse dans les méthodes d'enquête par échantillonnage.

Cette polémique ouvre l'ère de l'échantillonnage qui, dès le début, a fait l'objet de recherches assidues et fructueuses (notamment de la part de Jerzy Neyman, 1934, 1938, 1952). Celui-ci montre notamment qu'il est nécessaire de surreprésenter dans l'échantillon les catégories où la dispersion est la plus grande. Ce résultat rend obsolète le concept de représentativité encore trop souvent utilisé. Avec le développement des technologies de l'information, vient ensuite l'ère des registres et des fichiers administratifs dans les années 1970. Les données administratives sont souvent présentées comme la nouvelle source de données. Cependant, il ne faut pas oublier que certains pays comme la Finlande disposent d'un registre de population depuis plus de 50 ans. Les moyens techniques pour créer un registre existent depuis très longtemps. Les obstacles à l'utilisation des registres à des fins statistiques sont principalement d'ordre organisationnel, politique et juridique. Ils résultent souvent des règles de protection des données ou des difficultés de communication entre les différentes administrations. Depuis longtemps, les problèmes ne sont manifestement pas d'ordre technique.

Le développement de nouvelles sources de données n'a pas mis fin aux pratiques antérieures. De nombreux instituts nationaux de statistique effectuent encore des recensements et presque tous réalisent des enquêtes par sondage. L'intégration des fichiers administratifs dans la statistique officielle est loin d'être évidente. Les registres dépendent des spécificités administratives des pays, ce qui ne facilite pas l'harmonisation des statistiques. Les fichiers administratifs contiennent souvent de nombreuses erreurs, comme en témoigne la crise de Serafe en Suisse.

Depuis 2019, Serafe est une société privée chargée de collecter les redevances audiovisuelles en Suisse. Lors des premières factures envoyées, un grand nombre d'erreurs a été constaté. Serafe avait envoyé les factures sur la base des registres de "contrôle des habitants" mis à jour par les communes. De nombreuses erreurs semblaient être liées à la composition des ménages dans les immeubles à plusieurs logements ou à l'obsolescence des données. Ces registres de contrôle des habitants sont également utilisés à des fins administratives et statistiques, mais l'impact de ces erreurs est relativement faible pour ces applications. Dans les fichiers administratifs, on observe souvent que seules les variables nécessaires au fonctionnement administratif immédiat sont correctement mises à jour.

On ne peut donc pas s'attendre à ce que toutes la statistique officielle soit produite à partir du même type de source. Chaque type de source contient des erreurs spécifiques, qu'il s'agisse de données produites par des offices statistique ou non. Les problèmes à résoudre peuvent être les erreurs d'échantillonnage, les erreurs dues à la non-réponse, la sous-couverture, la sur-couverture, les doublons, les erreurs de mesure, les erreurs dues à la fraude. La statistique officielle est une lutte perpétuelle contre ces erreurs.

Dans chaque source, on peut trouver une fiabilité particulière. Un type de fiabilité est l'identification correcte des unités statistiques, qui peut souvent être obtenue grâce à un fichier ou un registre administratif. Cependant, ce registre peut contenir des variables de mauvaise qualité ou obsolètes. Une enquête par sondage, même avec de la non-réponse, peut contenir des mesures plus fiables pour les variables d'intérêt. Le problème se résume alors à combiner ce qui est le plus fiable dans les différentes sources. La question cruciale en matière de statistique officielle est donc l'intégration de données provenant de différentes sources afin d'améliorer au mieux la fiabilité de chaque source.

Dès le début, la recherche en statistique officielle s'est intéressée au problème de l'intégration des données. Il y a presque un siècle, la controverse entre, d'une part, Corrado Gini et Luigi Galvani et, d'autre part, Jerzy Neyman (Gini et Galvani, 1929) tourne autour de l'échantillonnage équilibré et de l'échantillonnage aléatoire. Gini et Galvani avaient sélectionné un échantillon de 29 districts (*circondari*) sur 214 de manière à rendre les mêmes moyennes que celles du recensement pour plusieurs variables connues. Il s'agit donc d'utiliser une source (le recensement) pour améliorer la collecte d'un échantillon. Neyman critique cette façon de faire, car la sélection de l'échantillon n'est pas aléatoire et il n'est donc pas possible de faire une inférence. Nous savons évidemment maintenant qu'il est possible de sélectionner un échantillon qui est à la fois équilibré et aléatoire (Deville et Tillé, 2004). Ken Brewer (2013) interprète cette discussion comme la deuxième controverse des méthodes d'enquête par échantillonnage.

Une autre question est l'ajustement des données d'enquête aux données de recensement.

A ce sujet, l'article de W. Edwards Deming et Frederick F. Stephan de 1940 est considéré comme un texte fondateur de la statistique officielle (Deming et Stephan, 1940). Le problème traité est l'ajustement d'un tableau obtenu par échantillonnage sur des totaux marginaux connus par un recensement. L'intégration de différentes sources est donc présente dès le début de l'utilisation des méthodes d'échantillonnage. Pour mémoire, l'article est cependant mathématiquement faux puisqu'il soutient que la méthode *raking ratio* s'obtient en minimisant la distance chi-carré sous les contraintes données par les totaux marginaux connus. Ceci n'est évidemment pas vrai puisque c'est la divergence de Kullback-Leibler qui doit être minimisée. Il n'en reste pas moins que cet article et tous ceux qui suivent cherchent à optimiser l'intégration des données d'enquête et de recensement.

L'article de Deville et Särndal (1992) qui définit la méthode générale de calage est l'aboutissement de cette recherche. Ces auteurs proposent une méthode de calage des données d'enquête sur les données de recensement en s'affranchissant de la modélisation. L'originalité est que le traitement est effectué de manière à obtenir un système de pondération applicable à n'importe quelle variable, ce qui rend son application extrêmement pratique. Deville et Särndal proposent également une méthode pour mesurer la précision des estimations obtenues. Cette méthode est devenue essentielle et typique de la statistique officielle. Elle est maintenant appliquée par tous les statisticiens d'enquête. (voir aussi Särndal, 2007; Devaud et Tillé, 2019a,b). L'utilisation du calage s'est généralisée, d'autant plus que le calage est aujourd'hui utilisé non seulement pour corriger l'erreur d'échantillonnage mais aussi pour corriger l'erreur de non-réponse et de mesure (voir entre autres Dupont, 1994; Fuller *et al.*, 1994; Lundström et Särndal, 1999; Deville, 2000; Särndal et Lundström, 2005; Kott, 2006; Brick, 2013; Valliant *et al.*, 2013; Haziza et Lesage, 2016; Devaud et Tillé, 2019a).

Un problème en développement est l'intégration non seulement de deux sources mais d'une multitude de sources distinctes. Les méthodes de calage peuvent être généralisées pour harmoniser plusieurs sources. Ces sources peuvent être deux échantillons ou plusieurs échantillons et un recensement : (voir entre autres Guandalini et Tillé, 2017). Yang et Kim (2020) donnent un aperçu des méthodes modernes d'intégration des données provenant de différentes sources.

L'utilisation des sources peut évoluer assez rapidement. De nombreux pays européens ont abandonné les grands recensements de population. Les statistiques basées sur ces recensements sont maintenant produites en utilisant des registres et de nouvelles enquêtes complètent ces registres. Les instituts nationaux de statistique réalisent encore beaucoup d'enquêtes par sondage. Les méthodes d'échantillonnage sont de plus en plus utilisées pour vérifier ou améliorer la qualité des recensements ou des registres. Ainsi, l'échantillonnage peut devenir davantage un outil de contrôle de la qualité qu'une méthode directe de production de données.

L'évolution des pratiques ne va pas sans soulever de nouvelles questions. La plus emblématique a été la démission de Martha Farnsworth Riche de la tête du United States Census Bureau. Selon le site Wikipedia contributors (2020) : "Bien qu'elle n'ait invoqué que des raisons personnelles pour justifier sa démission, celle-ci a été perçue comme un signe que les républicains du Congrès étaient en train de gagner dans leur combat pour empêcher le Census Bureau d'utiliser des techniques d'échantillonnage permettant de corriger le sous-

dénombrement persistant des minorités et d'autres groupes sous-représentés" [3]. Cette intervention politique dans la méthodologie est évidemment extrêmement préoccupante et va à l'encontre d'un principe d'indépendance méthodologique des Instituts nationaux de statistiques.

7 La modélisation dans la statistique officielle

Le débat sur la place du modèle dans la théorie de l'échantillonnage par sondage est interprété par Ken Brewer (2013) comme la troisième controverse de l'échantillonnage par sondage. L'idée d'introduire un modèle pour exploiter les informations auxiliaires résultant d'un recensement dans une enquête a été initialement proposée par ce même Ken Brewer (1963). Cependant, cette idée a été principalement développée par Royall (1970, 1971, 1976) (voir aussi Valliant *et al.*, 2000; Chambers et Clark, 2012). L'approche par le modèle consiste à construire un modèle qui est estimé à l'aide de l'échantillon, puis à prédire les données de la population qui ne font pas partie de l'échantillon.

L'approche basée sur le modèle s'oppose à l'approche basée sur le plan de sondage qui consiste à pondérer les unités de l'échantillon par l'inverse de leurs probabilités d'être sélectionnées. Les poids initiaux de l'enquête sont l'inverse des probabilités d'inclusion. Ces poids sont ensuite légèrement modifiés au moyen de la technique de calage de Deville et Särndal afin de restituer exactement les totaux des variables connues par un registre ou par un recensement.

Les statisticiens officiels sont par nature réticents à la modélisation. La raison en est la suivante : Modéliser, c'est en quelque sorte exprimer une opinion sur des observations. La modélisation peut donc être interprétée comme contraire au principe d'impartialité de la statistique officielle. Cependant, les approches basées sur le modèle et sur le plan ne sont pas nécessairement contradictoires. Särndal *et al.* (1992) préconise une approche basée sur la plan assistée par un modèle et néanmoins valable. Il est donc possible de construire le concept de double robustesse au sens d'une estimation qui serait valide soit lorsque le modèle est correct, soit lorsque les probabilités d'inclusion et de réponse sont correctement identifiées. Cette question est discutée entre autres dans Nedyalkova et Tillé (2008).

Cependant, il existe des problèmes qui ne peuvent être résolus sans une certaine modélisation. La recherche concernant l'estimation pour de petits domaines a été très active au cours des dernières décennies. Ces méthodes consistent à produire des estimations à des niveaux très bas (districts, communes) à partir d'une enquête par sondage et d'un registre ou d'un recensement. Une fois encore, le problème est d'utiliser au mieux les deux sources d'information. L'échantillon contient la variable d'intérêt. Le registre contient la liste de toutes les unités de la population et les variables auxiliaires. On utilise souvent des estimateurs composites, qui sont des mélanges d'estimations directes calculées à partir de l'échantillon et d'estimations obtenues à partir d'un modèle reliant les variables auxiliaires. Ces estimateurs composites

[3]Traduit de l'anglais "Although she cited only personal reasons in her resignation, it was seen as a sign that Congressional Republicans were winning in their fight to prevent the Census Bureau from using sampling techniques to correct for persistent undercounting of minorities and other underrepresented groups."

peuvent être obtenus au moyen de modèles mixtes pour lesquels les domaines sont des effets aléatoires : (voir le livre très complet de Rao et Molina, 2015).

Les instituts nationaux de statistique se sont toutefois montrés relativement prudents et ont rarement publié les estimations obtenues par ces méthodes. L'une des raisons est qu'il peut exister des particularités bien connues au niveau des communautés locales qui ne seraient pas prises en compte par un modèle. Par exemple, Molina et Strzalkowska-Kominiak (2020) ont proposé des estimations de la population active dans les districts suisses à l'aide de l'enquête structurelle suisse. Un district suisse limitrophe du Liechtenstein était particulier car un grand nombre de frontaliers suisses travaillent au Liechtenstein. Ce district a donc fait l'objet d'un traitement particulier. Cependant, s'il n'est pas possible d'identifier les singularités avant la modélisation, les estimations peuvent se révéler très éloignées de la réalité.

Un autre domaine où il est difficile de travailler sans modélisation est celui de la non-réponse. Il y a deux façons de procéder. Soit on peut prédire les valeurs manquantes (imputation), soit on peut estimer la probabilité que ces valeurs manquent afin de pondérer les observations qui répondent. Les modèles peuvent être très simples. On peut imputer par un plus proche voisin. On peut aussi imputer par un simple ratio ou par une prédiction par régression. On peut aussi prédire en prenant au hasard un individu appartenant à une petite strate homogène. Afin d'éviter une mauvaise spécification du modèle, on cherche souvent à traiter la non-réponse de manière doublement robuste. Deux modèles sont utilisés : le premier permet de prédire la valeur manquante et le second permet de prédire la probabilité d'être manquant. La double robustesse signifie que l'estimation est alors approximativement sans biais si au moins un des deux modèles est bien spécifié : (voir à ce sujet Kang et Schafer, 2007; Han et Wang, 2013; Kim et Haziza, 2014; Boistard *et al.*, 2016; Chen et Haziza, 2017).

8 Big Data et statistique officielle

Tim Harford (2014) donne un avis tranché sur le big data : "Comme beaucoup de mots à la mode, "big data" est un terme vague, souvent lancé par des gens qui ont quelque chose à vendre. Certains mettent l'accent sur l'ampleur des ensembles de données qui existent aujourd'hui - les ordinateurs du grand collisionneur de hadrons, par exemple, stockent 15 pétaoctets de données par an, soit l'équivalent d'environ 15 000 ans de votre musique préférée" [4]. En effet, on assiste souvent à des présentations d'"experts" annonçant l'ère des "big data" dont les diapositives ne contiennent que des listes de mots commençant par "V", des chiffres en yottabytes et des patates reliées par des flèches sans jamais donner une application réelle et concrète. Les diapositives sont agrémentées de citations pontifiantes et de dessins humoristiques. Faire des listes de mots commençant par une lettre est le contraire d'une démarche scientifique sérieuse.

[4]Traduit de l'anglais "As with so many buzzwords, "big data" is a vague term, often thrown around by people with something to sell. Some emphasise the sheer scale of the data sets that now exist – the Large Hadron Collider's computers, for example, store 15 petabytes a year of data, equivalent to about 15,000 years' worth of your favorite music."

La fascination pour le big data peut être irritante à plusieurs égards. Parler de "big data" dans la statistique officielle n'est pas approprié. Les grands fichiers administratifs tels que les registres de la population ou des entreprises ne peuvent pas être qualifiés de "big data". Ils peuvent contenir des millions d'enregistrements mais sont traitables sur n'importe quel ordinateur de bureau. Ces grands fichiers existent dans certains pays depuis plus de cinquante ans et ces pays ont toujours trouvé les moyens informatiques de les gérer. Ce ne sont pas des big data.

Je suis convaincu que l'utilisation du web scraping restera très limitée dans la production de statistiques officielles. Par exemple, l'inflation ne peut pas être calculée exclusivement en allant chercher automatiquement les prix sur Internet. ten Bosch *et al.* (2018) présentent une liste d'expériences et jettent les bases d'une méthodologie pour le web scraping dans la statistique officielle. Ils montrent que, dans la plupart des cas, les projets finissent rarement par être mis en production. Les principales raisons sont que la statistique publique doit être pérenne et qu'Internet n'est pas stable dans le temps. Ensuite, parce que la statistique officielle doit avoir une méthodologie et des sources documentées, parce qu'il ne suffit pas de regarder un prix sur Internet, il faut aussi pouvoir les vérifier. Et surtout parce que les instituts nationaux de statistique n'ont pas attendu la mode des big data pour utiliser directement les fichiers de prix des principaux distributeurs. Souvent, les données collectées en observant le web peuvent être plus simplement obtenues par un autre moyen, plus stable.

Le nombre d'applications du *web scrapping* réellement mises en pratique pour la production de statistiques officielles est donc très limité. On peut citer quelques applications dans les indices de prix pour des produits vendus presque exclusivement via Internet. Par exemple, le Bureau américain de statistique du travail (*Bureau of Labor Statistics*) utilise le web pour suivre l'évolution des prix des billets d'avion (US Bureau of Labor Statistics, 2021). Cependant, l'observation des prix sur le web ne peut pas vraiment être appelée "web scraping". L'analyse des réseaux sociaux peut toutefois être intéressante pour l'analyse des sciences sociales, comme le note Connelly *et al.* (2016). Néanmoins, la statistique officielle ont besoin d'un système de production stable et reproductible permettant des comparaisons temporelles, ce qui est difficile à obtenir à partir du web.

Il existe aussi le "*billion price project*" qui prévoit de calculer un indice des prix en utilisant le web scraping (voir Cavallo et Rigobon, 2016). Cette initiative ne vient pas du monde de la statistique officielle mais de deux professeurs du MIT Sloan et de la Harvard Business School. Cette approche peut être intéressante lorsque la statistique officielle est déficiente, comme ce fut le cas en Argentine, mais elle n'a pas été intégrée aux statistiques officielles jusqu'à présent. Je suis également convaincu que la méthodologie est plus importante que la taille de l'échantillon et que l'observation de plus d'un milliard de prix n'est pas nécessairement une garantie de la qualité des estimations.

Les énormes flux de données des réseaux sociaux n'ont pas non plus de valeur. Ils appartiennent aux géants du web. Ces données ont une immense valeur commerciale pour effectuer du profilage publicitaire. Il serait également extrêmement dangereux de devenir dépendant de ces géants. Surtout, ces données ne sont pas fiables en termes d'identification d'unités statistiques et de variables potentiellement utilisables.

Il existe cependant des domaines où les données sont vraiment massives, comme l'analyse

d'images pour les statistiques territoriales. Dans ce cas, les données peuvent être produites (photos aériennes ou satellites, cartographie) par le pays qui les utilise. Par exemple, le *Joint Research Center* de l'Union européenne à Ispra a une longue tradition d'estimation de la production agricole à partir d'images satellites (voir par exemple Gallego *et al.*, 1993; Taylor *et al.*, 1997; Gallego, 2004; Carfagna et Gallego, 2005; Kussul *et al.*, 2016). Cependant, cette pratique a été développée avant la popularisation du mot "big data". En effet, l'analyse d'images est un domaine de recherche qui a émergé avec les débuts de l'informatique. Des progrès significatifs ont été réalisés, notamment grâce à l'utilisation des réseaux de neurones, ce qui nous permet d'envisager un développement important des applications dans la statistique publique.

9 Apprentissage statistique en statistique officielle

Nous parlons souvent de nouvelles méthodes statistiques. La plupart de ces méthodes ne sont en fait pas si nouvelles et ont toutes été développées au cours du 20ème siècle. Plusieurs méthodes (forêts aléatoires, machine à vecteur de support, réseau de neurones, plus proches voisins) permettent de faire des prédictions sans avoir à réfléchir aux relations entre les variables dépendantes et la variable d'intérêt. Peut-on suivre la célèbre phrase de Deng Xiaoping qui disait : "Peu importe qu'un chat soit noir ou blanc, s'il attrape des souris c'est un bon chat" ? Peut-on utiliser dans la statistique publique des méthodes qui permettent de prédire sans comprendre ? En effet, la Charte statistique du Conseil d'éthique de la statistique publique suisse précise que "Les informations statistiques sont documentées afin d'en faciliter la compréhension et d'en permettre l'utilisation correcte". Est-ce compatible avec les méthodes d'apprentissage statistique ? Est-il suffisant de spécifier la méthode utilisée ?

Nous pensons que les méthodes d'apprentissage statistique peuvent être utilisées, mais avec certaines précautions. L'estimateur par régression généralisée (voir Särndal *et al.*, 1992) permet d'incorporer une prédiction dans l'estimation tout en restant approximativement sans biais sous le plan d'échantillonnage. Cet estimateur évite ainsi les dérapages dus à une mauvaise spécification du modèle. Les statisticiens officiels utilisent donc des modèles assez simples qui peuvent être intégrés dans une méthode qui reste valide sous le plan de sondage.

On pourrait penser que la statistique publique n'est pas concernée par la crise de la valeur p car elles produisent principalement des statistiques descriptives. Or, nous avons vu que sous cette production, se cache une machinerie très technique. Les enquêtes sont imputées, pondérées par des modèles de non-réponse, puis calées. Il faut donc choisir des variables de calage, concevoir des modèles de non-réponse et des modèles d'imputation. Ces opérations nécessitent une modélisation et cette modélisation est souvent basée sur des tests d'hypothèses, c'est-à-dire des valeurs p. L'utilisation des statistiques paramétriques est donc une tâche cachée. Dans ce domaine, l'utilisation de nouvelles méthodes d'apprentissage profond peut être très prometteuse.

Les statisticiens officiels connaissent également depuis longtemps des méthodes non-paramétriques simples pour traiter la non-réponse, comme l'imputation par le plus proche voisin ou par un individu sélectionné dans une strate homogène. Des méthodes comme le *support*

vector machine ou le *forêts aléatoires* consistent finalement à fractionner l'espace des variables explicatives pour définir un ou plusieurs voisins de l'unité afin de réaliser une prévision.

L'estimateur de régression généralisée a permis à Breidt et Opsomer (2000), par exemple, de construire des estimations assistées par un modèle dont les prédictions sont réalisées par la méthode des polynômes locaux. En suivant cette même approche, toutes les méthodes de prévision peuvent être utilisées sans introduire de risques disproportionnés dans les estimations. Ainsi, les travaux récents de Beaumont et Bocci (2008); Goga et Shehzad (2014); Breidt *et al.* (2017); McConville *et al.* (2017); Mayor-Gallego *et al.* (2019); Chen *et al.* (2019); Tan (2020); Dagdoug *et al.* (2020a,b) qui intègrent les méthodes de *shrinkage* et de *apprentissage statistique* dans le calage et le traitement de la non-réponse, montre probablement la voie à suivre pour les recherches futures en matière de statistique officielle.

Les méthodes dites *shrinkage* comme le Lasso permettent de choisir les variables dans une modélisation ou un calage. Les statisticiens d'enquête ont souvent tendance à sur-caler les enquêtes car de plus en plus de variables sont disponibles dans les registres. Or, le calage est avant tout une technique d'estimation qui vise à réduire les variances des estimations. En général, le calage comme la modélisation doivent être soumis à un principe de parcimonie. Il consiste à trouver le modèle le plus efficace possible tout en étant le plus simple possible. La méthode Lasso permet de réduire le nombre de variables de calage. Elle a déjà été appliquée aux données d'enquêtes par McConville *et al.* (2017).

10 Conclusions

Le terme big data devrait peut-être être abandonné. Lorsque l'on demande ce qu'est le "big data", on obtient souvent une réponse digne de la classification des animaux établie par Borges dans sa nouvelle "Langue analytique de John Wilkins" (Borges, 2012). Le terme "big data" peut désigner à la fois certains types de données ou un ensemble de méthodes. Par exemple, l'analyse d'images, l'analyse de flux de réseaux sociaux, les grands fichiers administratifs, les données de compteurs intelligents, les méthodes de réseaux neuronaux, les méthodes de forêts aléatoires, les méthodes de machines vectorielles de support, les méthodes de statistiques *sparse* (comme le Lasso). Le mot "big data" regroupe donc un ensemble hétérogène de données et de méthodes dans lequel il est nécessaire de faire une distinction. Il est important de garder les choses en perspective. L'analyse d'images est un problème très ancien qui est utilisé en statistique publique depuis au moins 40 ans. Les "nouvelles" méthodes statistiques ont presque toutes été développées au siècle dernier, avant la mode du big data. Évidemment, toutes les "nouvelles" méthodes statistiques sont intéressantes et sont de plus en plus appliquées dans les statistiques publiques, parfois à des problèmes "classiques" tels que le traitement de la non-réponse lorsque les ensembles de données ne sont pas nécessairement de grande taille. La valorisation des données administratives est un défi très important. Cependant, je suis particulièrement sceptique quant à la production de statistiques directement à partir du web.

La multiplication des sources dans la statistique officielle peut être trompeuse car elle n'implique pas nécessairement une amélioration de la qualité : (voir entre autres Deville,

1997). Comme nous le rappelle Tim Harford (2014), l'abondance de données n'est pas et ne sera jamais synonyme de qualité. Les fichiers administratifs contiennent souvent un grand nombre d'erreurs car ils n'ont pas été conçus pour être utilisés à des fins statistiques. Ces nouvelles sources doivent être combinées avec toutes les autres sources disponibles. L'intégration des sources est une question à laquelle les chercheurs se sont attaqués depuis les débuts de la statistique officielle. Ces méthodes doivent encore être développées, car de plus en plus de sources seront disponibles. Il serait utile de disposer d'un cadre théorique général pour l'intégration des données.

Les nouvelles méthodes statistiques doivent également être intégrées ou combinées aux méthodes existantes. Nous pensons qu'il n'y aura certainement pas de table rase de la méthodologie mais une évolution vers des pratiques plus variées et plus axées sur les types de données à traiter. Plus que jamais, la statistique publique doit promouvoir et développer la recherche méthodologique autour de ses problématiques.

Bibliographie

Bastin, C., Benzécri, J. P., Bougarit, C. et Cazès, P. (1980). *Pratique de l'Analyse des Données*. Dunod, Paris.

Beaumont, J.-F. et Bocci, C. (2008). Another look at ridge calibration. *Metron*, 66(1):5–20.

Bellhouse, D. R. (1988). A brief history of random sampling methods. *In* Krishnaiah, P. R. et Rao, C. R., éditeurs : *Handbook of Statistics Volume 6: Sampling*, pages 1–14, New York, Amsterdam. Elsevier/North-Holland.

Benzécri, J.-P. (1973a). *L'analyse des données : tome 1 : La taxinomie*. L'analyse des données. Bordas, Paris.

Benzécri, J.-P. (1973b). *L'analyse des données : tome 2 : L'analyse des correspondances*. L'analyse des données. Bordas, Paris.

Bethlehem, J. G. (2009). The rise of survey sampling. The Hague, Statistics Netherlands.

Boistard, H., Chauvet, G. et Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics*, 43(3):683–699.

Borges, J. (2012). *Inquisiciones — Otras inquisiciones*. Penguin Random House Grupo Editorial España.

Box, G. E. P. et Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses*, volume 649. John Wiley & Sons, Hoboken.

Breidt, F. J. et Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.

Breidt, F. J., Opsomer, J. D. *et al.* (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5:5–13.

Brewer, K. R. W. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39(2):249–262.

Brick, M. J. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3):329–353. cited By 32.

Capel, R., Monod, D. et Müller, J.-P. (1996). Essai sur le rôle des tests d'hypothèse en sciences humaines. *Actualités Pédagogiques*, 1:1–51.

Carfagna, E. et Gallego, F. J. (2005). Using remote sensing for agricultural statistics. *International statistical review*, 73(3):389–404.

Cavallo, A. et Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2):151–78.

Chambers, R. L. et Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford.

Chen, J. K. T., Valliant, R. L. et Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):657–681.

Chen, S. et Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104(2):439–453.

Connelly, R., Playford, C. J., Gayle, V. et Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social science research*, 59:1–12.

Costantini, D. et Galavotti, M. C. (1986). Induction and deduction in statistical analysis. *Erkenntnis*, 24:73–94.

Dagdoug, M., Goga, C. et Haziza, D. (2020a). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. arXiv 2007.06298.

Dagdoug, M., Goga, C. et Haziza, D. (2020b). Model-assisted estimation through random forests in finite population sampling. arXiv 2002.09736.

Deming, W. E. et Stephan, F. F. (1940). On a least square adjustment of sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427–444.

Devaud, D. et Tillé, Y. (2019a). Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *TEST*, 4:1033–1065.

Devaud, D. et Tillé, Y. (2019b). Rejoinder on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *TEST*, 28:1087–1091.

Deville, J.-C. (1997). *Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?* Document de travail – Institut national de la statistique et des études économiques. Insee.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. *In Compstat – Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*, pages 65–76, New York. Springer.

Deville, J.-C. et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Deville, J.-C. et Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Dupont, F. (1994). Calibration used as a nonresponse adjustment, studies in classification, data analysis, and knowledge organization. *In* Diday, E., éditeur : *New Approaches in Classification and Data Analysis*, pages 539–548. Springer-Verlag.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the royal statistical society*, 98(1):39–82.

Fraser, D. A. S. et Reid, N. (2016). Crisis in science? or crisis in statistics! mixed messages in statistics with impact on science. *Journal of Statistical Research*, 48(1):1–9.

Fuller, W. A., Loughin, M. M. et Baker, H. D. (1994). Regression weighting in the presence of nonresponse with application to the 19871988 nationwide food consumption survey. *Survey Methodology*, 20:75–85.

Gallego, F. J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, 25(15):3019–3047.

Gallego, F. J., Delincé, J. et Rueda, C. (1993). Crop area estimates through remote sensing: stability of the regression correction. *International Journal of Remote Sensing*, 14(18):3433–3445.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2(67-78):1999.

Gelman, A. et Loken, E. (2014). The statistical crisis in science: data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American scientist*, 102(6):460–466.

Gini, C. et Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica*, Series 6, 4:1–107.

Goga, C. et Shehzad, M. A. (2014). A note on partially penalized calibration. *Pakistan Journal of Statistics*, 30(4):429–438.

Guandalini, A. et Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review*, 85:250–269.

Han, P. et Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430.

Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2:180–190.

Hansen, M. H. et Madow, W. G. (1974). Some important events in the historical development of sample survey. *In* Owen, D. B., éditeur : *On the History of Statistics and Probability*, pages 75–102. Marcel Dekker, New York.

Harford, T. (2014). Big data: A big mistake? *Significance*, 11(5):14–19.

Harman, G. et Kulkarni, S. (2012). *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press, Cambridge, Massachusetts.

Haziza, D. et Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1):129–145.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22:359–380.

Kang, J. D. Y. et Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9:176–183.

Kiær, A. N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11:180–185.

Kiær, A. N. (1903). Sur les méthodes représentatives ou typologiques. *Bulletin de l'Institut International de Statistique*, 13:66–78.

Kiær, A. N. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14:119–134.

Kim, J. K. et Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1):375–394.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:133–142.

Kruskal, W. et Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review*, 48:169–195.

Kussul, N., Lemoine, G., Gallego, F. J., Skakun, S. V., Lavreniuk, M. et Shelestov, A. Y. (2016). Parcel-based crop classification in ukraine using landsat-8 data and sentinel-1a data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2500–2508.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249.

Lundström, S. et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15:305–327.

Mayor-Gallego, J., Moreno-Rebollo, J. et Jiménez-Gamero, M. (2019). Estimation of the finite population distribution function using a global penalized calibration method. *AStA Advances in Statistical Analysis*, 103(1):1–35.

McConville, K. S., Breidt, F. J., Lee, T. C. M. et Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2):131–158.

Molina, I. et Strzalkowska-Kominiak, E. (2020). Estimation of proportions in small areas: application to the labour force using the swiss census structural survey. *Journal of the Royal Statistical Society*, A183(1):281–310.

Nedyalkova, D. et Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Neyman, J. (1938). Contribution to the theory of sampling human population. *Journal of the American Statistical Association*, 33:101–116.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School, U. S. Department of Agriculture, Washington.

Neyman, J. (1957). "Inductive Behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique*, pages 7–22.

Popper, K. (2005). *The logic of scientific discovery*. Routledge, London.

Quételet, A. (1846). *Lettres à S. A. R. le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités appliquées aux sciences morales et politiques*. M. Hayez, Bruxelles.

Rao, J. N. K. et Molina, I. (2015). *Small Area Estimation*. Wiley, New York.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57:377–387.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. *In* Godambe, V. P. et Sprott, D. A., éditeurs : *Foundations of Statistical Inference*, pages 259–279, Toronto, Montréal. Holt, Rinehart et Winston.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–664.

Särndal, C.-E. (2007). The calibration approach un survey theory and practice. *Survey Methodology*, 33:99–119.

Särndal, C.-E. et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Särndal, C.-E., Swensson, B. et Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Statistics Canada (2017). Statistics Canada's Quality Assurance Framework. Documentation of the internet site of Statistics Canada, Statistics Canada, Ottawa.

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158.

Taylor, J., Sannier, C., Delincé, J. et Gallego, F. J. (1997). Regional crop inventories in europe assisted by remote sensing. *Synthesis Report, Office for Publications of the European Commission*.

ten Bosch, O., Windmeijer, D., van Delden, A. et van den Heuvel, G. (2018). Web scraping meets survey design: Combining forces. *In Big Data Meets Survey Science Conference, Barcelona, Spain*.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society*, B73(3):273–282.

Tillé, Y. (2020). *Sampling and Estimation From Finite Populations*. Wiley, Hoboken.

Tukey, J. W. (1977). *Exploratory Data Analysis*, volume 2. Addison-Wesley, Reading, MA.

US Bureau of Labor Statistics (2021). Consumer price index `https://www.bls.gov/cpi/factsheets/airline-fares.htm`. web site visited on 2020-01-08.

Valliant, R., Dever, J. A. et Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

Valliant, R., Dorfman, A. H. et Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Wasserstein, R. L. et Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Wikipedia contributors (2020). Martha Farnsworth Riche – Wikipedia, the free encyclopedia. [Online; accessed 24-August-2020].

Yang, S. et Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3:625–650.