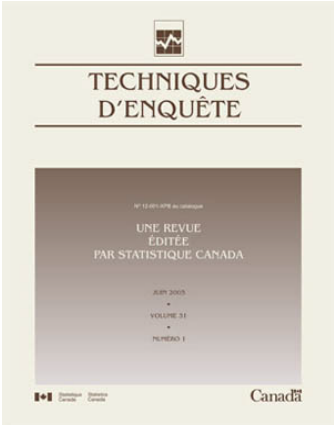
Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

# Commentaires à propos de l’article « Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle »




par Carl-Erik Särndal

Date de diffusion : le 3 janvier 2024

Statistique Canada Statistics Canada

Canada

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

**Courriel à** infostats@statcan.gc.ca

**Téléphone** entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

## Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2023
Vol. 49, nº 2, pp. 365-372
Statistique Canada, nº 12-001-X au catalogue


# Commentaires à propos de l'article « Les contributions de Jean-Claude Deville à la théorie des sondages et à la statistique officielle »

**Carl-Erik Särndal**[1]

Résumé

Au cours des dernières décennies, de nombreuses façons différentes d'utiliser l'information auxiliaire ont enrichi la théorie et la pratique de l'échantillonnage. Jean-Claude Deville a contribué de manière importante à ces progrès. Mes commentaires permettent de retracer certaines des étapes qui ont conduit à une théorie importante pour l'utilisation de l'information auxiliaire : l'estimation par calage.

**Mots-clés :** Information auxiliaire; estimation par régression généralisée; estimation par calage.

## Introduction

J'ai l'honneur de réfléchir à l'article d'Ardilly, Haziza, Lavallée et Tillé sur Jean-Claude Deville et à ses nombreuses contributions à la théorie et à la pratique des enquêtes.

J'appelle cet article Ardilly, Haziza, Lavallée et Tillé (AHLT). Il est en partie consacré à l'estimation du total de population finie selon la théorie du calage. C'est sur ce point que je me concentre dans le présent article. Plus de 30 ans se sont écoulés depuis la publication de l'article influent de Deville et Särndal (1992) sur le calage. Une suite importante lui a succédé un an plus tard, également publiée dans le Journal of the American Statistical Association (JASA).

Un précurseur – et un cas particulier – de l'estimateur par calage est [traduction] « l'estimateur bien connu par la régression linéaire généralisée », comme le formulent AHLT. Présenté dans leur formule (3.5), il est généralement connu sous le nom d'estimateur par la régression (linéaire) généralisée (ERG). Cela me donne l'occasion d'examiner la façon dont l'estimateur ERG a ouvert la voie à l'estimateur par calage. La forme de l'estimateur ERG a évolué au milieu et à la fin des années 1970 et dans les années 1980; on peut dire qu'il est à l'origine de l'estimateur par calage.

La construction de l'ERG est le fruit de « l'argument de prédiction » : prédire le mieux possible les valeurs des variables à l'étude pour les unités de population non observées. D'autre part, un courant un peu plus tardif de la littérature est axé sur « l'argument de pondération » : concevoir et calculer l'estimateur du total de la population au moyen d'une pondération appropriée des valeurs des variables à l'étude $y_k$ observées pour les unités $k$ dans l'échantillon probabiliste $s$; de préférence une pondération plus « riche en renseignements » que la simple pondération à probabilités inverses d'inclusion $d_k = 1/\pi_k$.

Mes commentaires dans le présent article reflètent ma compréhension, qui a pris forme sur plusieurs décennies, de l'estimation en présence d'information auxiliaire. Ils constituent un recueil incomplet d'une

1. Carl-Erik Särndal, professeur émérite, Statistics Sweden, Lilla Norregatan 12, SE 27135 Ystad, Suède. Courriel : carl.sarndal@telia.com.

période de perfectionnement au cours de laquelle Jean-Claude Deville a joué un rôle important. De nombreuses autres personnes ont apporté leur contribution et auraient dû être reconnues dans l'article.

## Rencontre

J'ai rencontré Jean-Claude pour la première fois dans les Alpes suisses. Pas en faisant du ski, mais lors de la traditionnelle « école d'hiver » annuelle de statistique qui se tient à la station Les Diablerets. Nous y avions tous deux été invités à prendre la parole en 1987. On m'avait dit à l'avance que j'allais rencontrer un statisticien français relativement nouveau dans le domaine de l'échantillonnage, qui s'était fait connaître grâce à de récentes contributions intéressantes.

J'y ai rencontré Jean-Claude en personne et en pensée. Nous avons établi une relation et avons eu une compréhension mutuelle dès le début. Diplômé de la prestigieuse École polytechnique, Jean-Claude s'appuyait sur un solide bagage mathématique et d'excellentes compétences. De plus, son travail à l'Institut national de la statistique et des études économiques (INSEE) l'a familiarisé avec les organismes statistiques nationaux et leurs efforts pour produire des statistiques nationales exactes.

Au fur et à mesure que j'apprenais à le connaître, il m'a semblé qu'une combinaison unique le caractérisait : une vision mathématique qu'il appliquait de manière parfois étonnante, tout en possédant une expérience claire en matière de production de statistiques nationales. Il était très sensible à la résolution de problèmes intégrés dans un environnement pratique plus grand, tel qu'un organisme statistique national.

Comme il me l'a expliqué, dans la trentaine déjà, il avait décidé de changer de domaine de recherche pour se consacrer à la science des enquêtes statistiques. Le concept de l'échantillonnage probabiliste l'intriguait : procéder à une sélection, en ayant des probabilités connues, parmi un ensemble fini d'unités identifiables. Sa curiosité a été stimulée par l'incroyable variété des méthodes utilisées pour y arriver et par la façon d'estimer, à l'aide de variables auxiliaires, les paramètres de population à partir de données tirées d'un échantillon probabiliste.

De sept ans son aîné, j'avais derrière moi une plus longue expérience du domaine. Il m'a raconté la façon dont il avait systématiquement « assimilé » le domaine de la science des enquêtes statistiques, avait étudié la littérature importante, des livres classiques sur l'échantillonnage des années 1950 jusqu'aux travaux récents, dans la vigueur naissante que la théorie de l'échantillonnage a trouvée depuis 1970 environ. Notamment, il admirait le travail du théoricien tchèque de l'échantillonnage Jaroslav Hàjek. Il avait étudié le livre de Wiley de 1977 que j'ai cosigné, *Foundations of Inference in Survey Sampling* (les fondements de l'inférence dans l'échantillonage en français). Dans les années qui ont suivi, il est venu plusieurs fois au Canada et nous avons travaillé ensemble à l'Université de Montréal.

## Théorie de la randomisation

Je crois que la préférence idéologique de Jean-Claude était en phase, essentiellement, avec la théorie de la randomisation, le volet de la science des enquêtes statistiques fondé sur le plan. Il s'agit d'une approche dans laquelle l'inférence concernant la population finie est fondée sur le plan d'échantillonnage avec probabilités, en utilisant les probabilités d'inclusion connues des unités échantillonnées.

Ce positionnement n'était pas du tout évident pour quelqu'un qui apprenait le domaine à l'époque. Les années 1970 avaient apporté un flot d'idées, parfois contradictoires. La théorie fondée sur le modèle, selon laquelle l'inférence doit être plutôt fondée sur la structure stochastique énoncée dans le modèle hypothétique, a constitué un prétendant de taille à la théorie traditionnelle de la randomisation fondée sur le plan.

Il me semble donc que lorsque Jean-Claude s'est lancé dans l'apprentissage du domaine, il a pu se sentir attiré par le mouvement qui prône la théorie fondée sur un modèle et a pu en devenir un adepte convaincu. Mais cela n'a pas été pas le cas.

## Information auxiliaire

L'estimation fondée sur des idées concernant les renseignements supplémentaires et les valeurs $y$ prédites pour les unités non observées semble avoir été étudiée, ou mise à l'essai, dès les années 1950, dans des endroits comme les organismes statistiques nationaux, où se trouvaient des méthodologistes d'enquête perspicaces. À l'INSEE, en France, P. Thionet et Y. Lemel ont été parmi les personnes dont les travaux importants ont influencé Jean-Claude.

Un exemple important d'information auxiliaire est lorsque les valeurs sont connues pour toutes les unités de population pour une ou plusieurs variables, les variables $x$, considérées comme liées à la variable de l'enquête, la variable $y$. Les pays scandinaves, dont les registres de population sont diversifiés et de grande qualité, ont servi de terrain d'essai pour cette méthodologie. L'utilisation d'information auxiliaire a donné lieu, à partir des années 1970, à de grands progrès dans la recherche et la pratique en matière de statistiques d'enquête.

## Estimateur par la régression généralisée

L'estimateur ERG repose sur « l'argument de prédiction »; les valeurs $y$ pour les unités non échantillonnées sont prédites en fonction d'une relation perçue entre la variable à l'étude y et une ou plusieurs variables auxiliaires corrélées. L'estimateur ERG linéaire, tel que nous le connaissons aujourd'hui, donné dans la formule (3.5) d'AHLT, a évolué progressivement entre le milieu et la fin des années 1970. Plusieurs personnes y ont pris part; un raffinement progressif s'est opéré. La première fois que j'ai utilisé l'« estimateur par la régression généralisée » en tant que terme et principe de construction incomplet à l'époque, c'était, autant que je m'en souvienne aujourd'hui, dans un article cosigné par Cassel, Särndal et Wretman (1976) dans Biometrika. Parmi les contributions à « l'approche de la régression », les travaux menés à Iowa State par W.A. Fuller et ses étudiants se distinguent.

Le terme « généralisé » signifiait essentiellement que la première forme de l'estimateur ERG a permis d'élargir ce que Cochran et d'autres textes des années 1950 avaient présenté comme méthode de réduction de la variance, en attachant un terme d'ajustement par régression linéaire à l'estimateur de base sans biais par rapport au plan. L'ajustement, de faible ampleur dans les grands échantillons, était une fonction des variables auxiliaires.

De façon plus générale, supposons que $\hat{y}_k$ est la valeur prédite, au moyen d'un « modèle d'assistance » linéaire ou non linéaire, de la valeur de la variable étudiée $y_k$. La formulation ERG est $\hat{Y}_{\text{ERG}} = \sum_s d_k y_k + \left(\sum_U \hat{y}_k - \sum_s d_k \hat{y}_k\right)$, où $U$ représente la population et $s$ représente l'échantillon probabiliste de $U$. Les résidus $y_k - \hat{y}_k$ laissés par l'ajustement du modèle deviennent plus évidents lorsque nous l'écrivons comme suit $\hat{Y}_{\text{ERG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)$.

Dans des articles datant de la fin des années 1970 et du début des années 1980, j'ai eu l'occasion d'examiner l'estimateur ERG linéaire, c'est-à-dire lorsque $\hat{y}_k$ est le résultat d'un ajustement par régression linéaire. Je me souviens de l'article de Biometrika, cosigné en 1976, surtout en raison d'une curiosité administrative.

Bien que cela ait été exceptionnel pour une soumission courante à Biometrika, le rédacteur en chef, D.R. Cox, souhaitait accompagner l'article d'une discussion spéciale, ce dont nous avons convenu. L'article a été perçu, apparemment, comme le mélange peu orthodoxe d'un intérêt envers un modèle et ses propriétés et, en même temps, envers la distribution de randomisation, $p$, découlant de l'échantillonnage probabiliste.

L'animateur de la discussion était T.M.F. (Fred) Smith, un théoricien d'enquête hautement respecté avec qui j'ai eu le plaisir d'avoir, au fil du temps, de nombreux contacts fructueux et amicaux. Il a écrit : [traduction] « Cet article soulève une question fondamentale concernant l'inférence d'une population finie au moyen d'une superpopulation. […] Les auteurs imposent la contrainte supplémentaire que (l'estimateur par la régression) $T$ doit être non biaisé $p$ […]. Pourquoi les probabilités de sélection, $p$, devraient-elles avoir la priorité sur le modèle $\xi$ ? » Il s'agissait d'une question bien motivée de son point de vue fondé sur le modèle : si le modèle de régression était digne de confiance pour construire l'estimateur, pourquoi l'abandonner en faveur de la distribution de randomisation lorsqu'il s'agit d'évaluer les propriétés de base, telles que le biais et la variance ?

C'était une illustration d'un échange qui pouvait avoir lieu à l'époque entre un membre du nouveau mouvement fondé sur un modèle et un membre du mouvement traditionnel fondé sur le plan. En temps utile, la perspective de cet article s'est transformée en « estimation fondée sur le plan assistée par un modèle », c'est-à-dire assistée par le modèle mais ne dépendant pas de sa « véracité ». L'absence de biais $p$ asymptotique a servi de protection contre un modèle potentiellement faux ou inadéquat. La « véracité » ou non du modèle n'était pas la question principale. Si le modèle ne tient pas, l'estimateur est toujours asymptotiquement conforme au plan de sondage.

## Raisonnement réorienté

La littérature des années 1980 et des années suivantes propose une « réorientation du raisonnement »; une partie de l'attention a été détournée de l'argument de prédiction au profit d'un autre argument centré sur les poids attribués aux valeurs $y$ observées de l'échantillon.

Dans l'approche de prédiction, d'autres et moi avons construit un estimateur en prédisant les valeurs $y$ non observées aussi bien que possible, au moyen d'une sorte d'ajustement par la régression, et en utilisant les variables auxiliaires. Dans l'intérêt d'avoir une estimation précise, il était clair que les prédictions $\hat{y}_k$

devaient refléter aussi précisément que possible la valeur inconnue $y_k$ pour les unités non observées; les résidus $\hat{y}_k - y_k$ devaient être petits.

L'argument de pondération met plutôt l'accent sur l'estimation au moyen d'une pondération appropriée des valeurs $y$ observées dans l'échantillon. Dans l'intérêt d'avoir une estimation précise, le poids accordé à une unité échantillonnée doit refléter ce qui est connu et ce qui est particulier à propos d'une unité de la population. Les « bons poids » pourraient ou devraient être des fonctions dépendantes de l'échantillon des valeurs du vecteur auxiliaire $\mathbf{x}_k$. Ils pourraient n'impliquer qu'un petit, mais important, ajustement des poids de sondage de base $d_k = 1/\pi_k$, comme c'est le cas pour les poids sous-entendus par la formule de l'estimateur ERG.

Dans Särndal (1982), j'ai écrit l'estimateur ERG linéaire comme une somme pondérée selon l'échantillon, la valeur de la variable étudiée $y_k$ observée recevant le poids $d_k g_k$, où le poids du plan d'échantillonnage $d_k = 1/\pi_k$ subit un petit ajustement $g_k$, légèrement éloigné de « un ». Bethlehem et Keller (1987) montrent également l'interprétation de la pondération linéaire de l'estimateur ERG et constatent la propriété de calage des poids.

## Échantillonnage assisté par un modèle

L'épais manuscrit du livre de Särndal, Swensson et Wretman portant ce titre (Model Assisted Survey Sampling) était sur le point d'être achevé lorsque Jean-Claude et moi avons collaboré à la fin des années 1980. Je lui ai décrit l'esprit du livre. Publié en 1992 par Springer-Verlag, il est devenu populairement connu sous le nom de « Livre jaune ». Il préconise une perspective de l'inférence fondée sur le plan, plus particulièrement dans une forme communément appelée « assistée par un modèle ». Le rôle du modèle a été longuement expliqué, notamment au chapitre 6; la formulation était importante; elle était convaincante et a inspiré les autres.

Aussi curieux que cela puisse sembler aujourd'hui, en tant qu'auteurs de livres, nous étions très sensibles au climat de la théorie des enquêtes statistiques. Depuis le début des années 1970, la question de savoir si l'on peut ou non s'appuyer sur des modèles avait fait l'objet de nombreux débats dans les ouvrages consacrés à la théorie de l'échantillonnage d'enquête. Nous avions eu notre part de frustrations, dans le cas avec l'article susmentionné de Biometrika de 1976 sur l'estimation par régression généralisée.

Le Livre jaune a utilisé l'argument de la prédiction pour construire et expliquer avec soin l'estimateur ERG de $Y = \sum_U y_k$. Les valeurs prédites de $\hat{y}_k$ à partir d'un ajustement de la régression linéaire sont données par $\hat{y}_{ks} = \mathbf{x}_k' \mathbf{B}_s = \mathbf{x}_k' \left( \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left( \sum_s d_k \mathbf{x}_k y_k \right)$, où $s$ est l'échantillon probabiliste et $d_k = 1/\pi_k$. Elles sont calculables pour toutes les unités si $\mathbf{x}_k$ est disponible pour toutes. La forme linéairement pondérée de l'estimateur ERG est également soulignée :

$$\hat{Y}_{\text{ERG}} = \sum_s d_k g_k y_k,$$

où $g_k = 1 + \left( \sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)' \left( \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k.$

## Les poids *g*

Le poids *g*, comme nous avions l'habitude d'appeler $g_k$, est un facteur de pondération, égal à un plus un terme d'importance mineure dans les grands échantillons, mais dont l'impact est important. Il modifie légèrement le poids de sondage $d_k = 1/\pi_k$ en un poids total $d_k g_k$, et, comme l'explique le Livre jaune, ces poids satisfont

$$\sum_s d_k g_k \mathbf{x}_k = \sum_U \mathbf{x}_k .$$

C'est ce qu'on a appelé la propriété de calage. Je me souviens très bien de mon étonnement momentané à la suite de ma propre « découverte » de la propriété; c'était toutefois tout à fait évident après un examen plus approfondi et ça n'a donc pas fait l'objet d'une attention immédiate. D'autres personnes actives dans le domaine en étaient sans doute également conscientes vers les années 1980. Les poids $d_k g_k$, leur fonction, leur propriété de calage et leur utilisation dans l'estimation de la variance ont été au centre de l'attention dans un article de Biometrika par Särndal, Swensson et Wretman (1989).

La propriété de calage des poids $d_k g_k$ de l'ERG a cependant été un signal de mon intérêt pour les systèmes de poids calés : il doit s'agir d'une idée plus fructueuse en général. Lorsque Jean-Claude et moi en avons discuté à la fin des années 1980, la propriété était un fait bien établi, un point de départ : il doit être possible de l'étendre et de la généraliser.

## L'essor de la théorie du calage

Une phrase dans AHLT attire mon attention : les auteurs soutiennent que la théorie du calage, comme dans Deville et Särndal (1992), a apporté [traduction] « […] l'une des avancées les plus importantes dans le domaine de l'estimation en présence d'information auxiliaire : il est possible de construire l'estimateur ERG au moyen d'un calage ». En d'autres termes, une « découverte » a été que le champ d'application de la théorie du calage est suffisamment large pour admettre l'important estimateur ERG sous son ombrelle.

Il est en effet important que l'estimateur ERG fasse partie de la famille du calage; j'ajoute cependant que ma propre réflexion a eu lieu dans un ordre temporel différent : sachant à l'avance que les poids $d_k g_k$ de l'ERG ont la propriété de calage, il doit être possible d'étendre l'idée.

Cela a pris forme dans la théorie bien décrite dans AHLT, avec une mesure de la distance à minimiser entre les poids d'enquête $d_k = 1/\pi_k$ et les nouveaux poids $w_k$, sous la contrainte de calage $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Un poids calé résultant $w_k$ est une fonction de la valeur $\mathbf{x}_k$ du vecteur auxiliaire.

## La théorie du calage : implications et interprétations

J'ai commencé les présents commentaires en notant deux façons de procéder, « l'argument de la prédiction » par opposition à « l'argument de la pondération ». Cette distinction importante a marqué une partie du développement théorique de l'échantillonnage d'enquête au cours des dernières décennies.

Le premier a été utilisé avec succès pour créer l'estimateur ERG (linéaire ou non linéaire). Il me semble maintenant que l'argument de la pondération, tel qu'il est utilisé dans la théorie du calage, a une plus vaste application, ou un attrait plus grand. Pour faire une estimation, il suffit d'additionner les valeurs $y$ correctement pondérées.

Depuis l'article de Deville et Särndal (1992), beaucoup de choses ont été dites sur le calage dans la littérature. Il est vrai que le calage y était présenté comme une méthodologie fondée sur le plan, mais dans des conditions d'enquête idéales, avec un taux de réponse de 100 % et l'absence d'autres erreurs d'enquête. Cependant, avec le temps, le concept de calage s'est révélé être un instrument d'une puissance et d'une flexibilité extraordinaires, comme en témoigne, par exemple, la variété appelée calage basé sur un modèle, et en particulier ses extensions à l'ajustement pour la non-réponse, qui font l'objet d'une vaste littérature à part entière.

### Une image

Je termine par une digression sur le mot « calage ». À l'INSEE, le terme « calage » était apparemment bien établi depuis longtemps. Il a été utilisé, semble-t-il, dans les premiers exemples de pondération qui confirment le total connu de la population d'une variable auxiliaire.

En français, le verbe « caler » signifie immobiliser, fixer, stabiliser. Ailleurs aussi, il y avait sans doute des statisticiens perspicaces qui dérivaient et calculaient des poids, avec la propriété aujourd'hui appelée « calée », bien avant qu'il n'y ait un nom spécial et une théorie spéciale pour la procédure.

Pour l'article du JASA de 1992, Jean-Claude et moi avons opté pour le terme « calage », conscients du sens qu'il a pour certains, notamment dans l'idée de la balance, cet instrument utilisé dans les magasins d'alimentation il y a bien longtemps : deux plateaux à chaque extrémité d'un fléau. Pour répondre à la commande d'un client, disons 600 grammes de café, de beurre ou de farine, l'employé du magasin place des poids sur un plateau, pour un total de 600 grammes, puis mesure l'aliment souhaité sur le plateau opposé, jusqu'à ce que l'équilibre soit atteint. Le magasin disposait d'une collection de poids en métal, de différentes dénominations; ils étaient certifiés, calés, afin de garantir le droit du client au poids correct. Comme certains me l'ont rappelé avec tendresse au fil des ans, l'expression « poids calés » évoque cette vieille image mentale d'une procédure digne de confiance.

## Bibliographie

Bethlehem, J.G., et Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

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

Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

Särndal, C.-E., Swensson, B. et Wretman J.H. (1989). The weighted residual technique for estimating the variance of the finite population total. *Biometrika*, 76, 527-537.

Särndal, C.-E., Swensson, B. et Wretman J.H. (1992). Model assisted survey sampling. New York: Springer-Verlag.