STATISTIQUE ENVIRONNEMENTALE, VARIABLES NON CORRÉLÉES : ENJEUX DE L'INTÉGRATION DES ENQUÊTES ENVIRONNEMENTALES AU PROGRAMME INTÉGRÉ DE LA STATISTIQUE DES ENTREPRISES (PISE) DE STATISTIQUE CANADA

Herbert Nkwimi-Tchahou[1] & Martin Hamel[2]

[1] *Statistique Canada: Division des Méthodes d'enquêtes auprès des entreprises. 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6. Immeuble R.-H.-Coats Étage 22 O*
Herbert.nkwimitchahou@canada.ca
[2] *Statistique Canada: Division des Méthodes d'enquêtes auprès des entreprises. 100, promenade Tunney's Pasture, Ottawa ON K1A 0T6. Immeuble R.-H.-Coats Étage 22 F*
Martin.hamel@canada.ca

Résumé. Depuis 2010, Statistique Canada a commencé le développement du programme intégré de la statistique des entreprises (PISE) visant à maximiser le processus de production de données statistiques pour les enquêtes entreprises grâce à la normalisation et à l'utilisation de services organisationnels et de systèmes généralisés. L'intégration des enquêtes environnementales dans ce cadre normalisé des enquêtes économiques soulève de nombreux défis à cause notamment du caractère spécifique des données collectées. Les données environnementales d'une entreprise sont en général très peu reliées aux variables économiques standards comme par exemple le revenu, le nombre d'employés ou les dépenses. Dans cet article, nous présenterons quelques enjeux rencontrés ainsi que des solutions envisagées afin de maintenir le niveau de qualité habituel des estimations.

Mots-clés. Enquêtes environnementales, plan de sondage avec seuil d'exclusion, information auxiliaire

1 Introduction et contexte

Le programme intégré de la statistique des entreprises (PISE) à Statistique Canada a vu le jour lors de l'importante restructuration entreprise par l'Agence depuis 2010 et connue sous le terme d'Architecture Opérationnelle du Bureau (AOB). Les processus opérationnels, les règles administratives, les systèmes informatiques ainsi que l'architecture organisationnelle ont été optimisés, permettant ainsi de réduire les coûts en garantissant les plus hautes normes de qualité et d'actualité dans la prestation de ses services (voir Lahaie M. et al).

Le volet PISE de ce vaste projet, mis en place depuis avril de la même année vise à maximiser le processus de production de données statistiques pour les enquêtes entreprises grâce à la normalisation et à l'utilisation de services organisationnels et de systèmes généralisés. La plateforme PISE repose sur un ensemble de cinq modules. Le module de l'échantillonnage (SGECH), la vérification et l'imputation (BANFF), l'estimation (SGE), la confidentialité (G-CONFID) et la diffusion (G-EXPORT). Cet important projet vise six grands objectifs : 1) Améliorer la qualité des données (méthodologie et processus harmonisés, questionnaire harmonisé etc…) ; 2) Réduire le fardeau de réponse (utilisation maximale des données administratives, collecte active fondée sur les indicateurs de qualité et mesure d'impact) ; 3) Moderniser les infrastructures de traitement de données (modèle fondé sur les métadonnées) ; 4) Simplifier et normaliser les processus (réduction de la courbe d'apprentissage et des délais pour accroître la capacité de répondre aux nouveaux besoins statistiques) ; 5) Intégrer les enquêtes (intégrer un maximum d'enquêtes économiques) et enfin 6)

réduire les coûts (voir Yukman Cheung et al. pour plus de détails).
L'intégration des enquêtes environnementales dans ce cadre normalisé pour les enquêtes économiques soulève de nombreux défis à différentes étapes d'une enquête à cause du caractère spécifique des données collectées. Cet article fera un survol des principaux enjeux méthodologiques ainsi que les solutions envisagées en vue de garantir le niveau de qualité habituel des estimations produites. La suite du document s'organise comme suit : la section 2, présentera les enquêtes environnementales touchées par cette intégration au PISE, puis nous présenterons de manière détaillée, quelques enjeux soulevés en s'appuyant sur une des plus importantes enquêtes environnementale avant de clore avec une brève conclusion.

2 Enquêtes environnementales et enjeux soulevés par le PISE

Depuis l'avènement du plan vert du Canada pour un environnement sain, lancé par le gouvernement fédéral, les statistiques sur l'environnement sont devenues une priorité pour de nombreuses agences gouvernementales. On a assisté à une croissance des enquêtes en lien avec l'environnement. À ce jour, la Division des méthodes d'enquêtes auprès des entreprises de Statistique Canada traite de nombreuses enquêtes environnementales parmi lesquelles on retrouve: l'enquête sur les usines de traitement de l'eau potable (EUTEP), l'enquête sur les dépenses de protection de l'environnement (EDPE), l'enquête sur l'eau dans les industries (EEI), l'enquête sur les biens et les services environnementaux (EBSE). Il s'avère important de regarder si les ajustements à la méthodologie sont nécessaires pour ces enquêtes afin de faciliter leur intégration dans le PISE. La présente section est consacrée à quelques-uns des obstacles rencontrés. Nous nous servirons de l'enquête sur l'eau dans les industries pour illustrer quelques défis. Nous commencerons par présenter brièvement cette enquête et ses objectifs, suivi des enjeux et quelques solutions explorées.

2.1 Enquêtes sur l'eau dans les industries (EEI)

L'enquête sur l'eau dans les industries est une enquête qui a lieu tous les deux ans et qui collecte les données sur la quantité d'eau utilisée, les coûts, les sources d'approvisionnement ainsi que le traitement et l'évacuation de l'eau. Cette enquête est constituée de trois composantes indépendantes mais très similaires, ciblant chacune un secteur industriel en particulier. La première composante est un recensement qui cible les quelques 125 emplacements de centrales thermoélectriques au Canada. La seconde composante cible environ 800 emplacements d'extractions minières. Un échantillon aléatoire simple stratifié est appliqué pour tirer un échantillon d'environ 380 emplacements. Le tableau ci-dessous donne un aperçu.

Tableau1: Composantes de l'enquête sur l'eau dans les industries (année de référence 2011).

Composantes / taille population	Plan d'échantillonnage	Taille de l'échantillon
Centrale thermoélectriques N ≈ 125 emplacements	Recensement	n ≈ 125 emplacements
Extraction minières N ≈ 90 000 emplacements	STR	n ≈ 380 emplacements
Manufacturier N ≈ 90 000 emplacements	STR avec seuil d'exclusion Royace-Maranda (R-M) à 5%	n ≈ 5000 emplacements

2.2 Enjeux soulevés par l'intégration de l'EEI au PISE

2.2.1 Combiner trois composantes ? Entreprise comme unité d'échantillonnage ?

Comme nous l'avons déjà mentionné à l'introduction, l'un des objectifs majeur du PISE est l'optimisation du processus de traitement des enquêtes. Compte tenu de la similarité entre les trois

composantes d'EEI, il est assez naturel d'envisager le traitement de celles-ci dans le PISE comme une seule enquête. Cette stratégie vise à réaliser un gain en temps et en ressources. Il convient de rappeler que la production de ces trois composantes se fait actuellement de manière indépendante; c'est à dire : chaque composante possède son système d'échantillonnage, de vérification et d'imputation, d'estimation ainsi que celui portant sur la confidentialité. L'idée de combiner ces trois composantes, bien qu'attrayante en apparence soulève tout de même un certain nombre de préoccupations.

2.2.1.1 Gestion efficace des ressources et de logistique

Le mode de traitement actuel, de la création de la base de sondage à la sélection de l'échantillon, est un processus linéaire en ce sens qu'on achève une activité relative à une composante avant d'entreprendre la suivante. Le traitement simultané des trois composantes nécessitera une planification minutieuse et rigoureuse afin de respecter les délais.

2.2.1.2 Augmentation du fardeau de réponse

Échantillonner au niveau «entreprise» peut entrainer une augmentation du fardeau de réponse. Contrairement à de nombreuses enquêtes économiques, l'emplacement est le lieu approprié pour la collecte des données en lien avec l'environnement. Un questionnaire envoyé à une entreprise sera réacheminée à ses emplacements dans la plupart des cas pour obtenir l'information demandée. Il existe des entreprises manufacturières ayant de nombreux emplacements (au-delà de 100 par exemple). Si une telle entreprise est sélectionnée, on sélectionnera la grappe entière de tous ses emplacements. Ce qui constitue un fardeau et peut augmenter la non réponse. En 2013 par exemple, la base de sondage contenait environ 123 400 emplacements correspondant à 119 000 entreprises. Une trentaine d'entreprises détenaient 20 emplacements et plus chacune. Pour ces dernière, pas moins de 10% d'emplacements étaient sous le seuil d'exclusion c'est à dire étaient classifiés dans la strate à tirage nul. Sélectionner de telles entreprises imposera un fardeau assez lourd à ces petits emplacements.

2.2.1.3 Stratification

En principe, dans un plan d'échantillonnage stratifié, chaque unité d'échantillonnage appartient à une et une seule strate. Or certaines entreprises ont des emplacements dans plus d'un des trois secteurs industriels ciblés. Affecter une entreprise à un et un seul secteur a des répercussions sur la couverture des deux autres. Par exemple, la consommation d'eau pour une entreprise donnée peut être négligeable pour le secteur manufacturier mais énorme pour le secteur minier. Si une telle entreprise est affectée au secteur manufacturier, il y a des chances qu'elle soit classée dans la strate à tirage nul, si on se base sur le revenu avec lequel cette dernière n'est pas corrélée. Son exclusion de l'échantillonnage peut induire un biais important pour le secteur minier. De manière plus générale, le biais associé à la couverture peut être considérable dans certains scénarios d'affectation.

Nous avons mené, conjointement avec les spécialistes du sujet, de nombreuses analyses et avons évalué différents scénarios. Il a été convenu que les trois composantes seront combinées dans le PISE mais l'unité d'échantillonnage sera l'emplacement et non l'entreprise.

2.2.2 Biais induit par la strate à tirage nul

Au-delà de la difficulté supplémentaire qu'apporte le caractère environnemental de nos enquêtes, produire des estimations non biaisées en présence de strate à tirage nul est un défi qui demeure d'actualité. Utiliser les estimateurs traditionnels naïfs peut conduire à des estimations fortement biaisées surtout si la frange de la population non enquêtée diffère du reste par rapport à la variable d'intérêt. Cette problématique a fait l'objet de nombreux articles récents. Par exemple, Haziza et al. (2010) montre qu'en présence de variable auxiliaire reliée aussi bien à la variable d'intérêt qu'à la probabilité d'exclusion, on peut réduire le biais dû à la couverture.

Dans la mise en place du PISE, il a été suggéré l'utilisation de l'algorithme de Royce-Maranda pour la création de strate à tirage nul, avec un seuil d'exclusion de 10%, basé sur une variable auxiliaire

pour toutes les enquêtes. Il s'agit pour EEI d'envisager l'utilisation de seuil d'exclusion pour les composantes centrale thermoélectrique et minière, ainsi qu'une augmentation du seuil de 5 à 10% pour la composante manufacturière. Il est important de rappeler qu'il n'existe pas sur la base de sondage de variables auxiliaires fortement reliées aux variables d'intérêts. Nous avons mené des analyses supplémentaires afin de cerner la qualité des estimations avec un tel scénario. Dans la sous-section suivante, après avoir présenté un bref rappel sur les plans avec seuil, nous présenterons quelques résultats observés.

2.2.2.1 Rappel sur les plans avec seuil d'exclusion

Soit U une population de taille N pour laquelle nous sommes intéressés à mesurer le total t_y d'une caractéristique y. On peut partitionner la population en deux strates de sorte que : $U = U_E \cup U_S$ et $N = N_E + N_S$. Un échantillon s est sélectionné dans U_S et les unités de U_E sont exclues de la population. L'objectif est de produire une estimation $\hat{t}_y$ de t_y total de y pour la population U à partir de l'échantillon. Si on utilise l'estimateur $\hat{t}_y = \sum_{k\in s} d_k\ y_k$, où les d_k sont des poids de sondage, on peut produire une estimation sans biais du total t_S de la sous population U_S. Afin de produire une estimation pour toute la population, il est courant de supposer que $t_E = \delta\ t_S$. Le facteur δ est généralement inconnu et peut être estimé par :

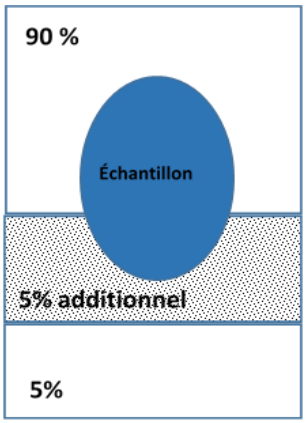
$$\hat{\delta} = \frac{\sum_{k\in U_E} X_k}{\sum_{k\in U} X_k}$$

X étant une variable auxiliaire disponible pour toute la population. On peut montrer que le biais $b(\hat{t}_y) = (\hat{\delta} - \delta)t_S$, voir Bee, M. et al. (2007). Il apparaît clairement que la source du biais est l'incapacité à produire un estimateur précis de δ et $|\hat{\delta} - \delta|$ renseigne sur l'ampleur du biais. Ce biais sera nul si la variable d'intérêt est proportionnelle à la variable auxiliaire. Ce qui n'est pas le cas de la variable revenu, utilisée pour définir le seuil d'exclusion dans notre enquête.

2.2.2.2 Quelques résultats

Dans cette partie, nous présentons des résultats de quelques analyses effectuées sur les trois composantes. Dans une première analyse, les données collectées en 2011 et 2013 ont servi à explorer l'effet du passage d'un seuil de 5 à 10% dans l'algorithme de Royce-Maranda pour la composante manufacturière. Les estimations publiées pour ces années de référence utilisent une base de sondage pour laquelle on a appliqué un seuil de 5%. Pour mener notre analyse, nous avons identifié les unités à exclure si on appliquait un 5% supplémentaire et avons évalué grâce à l'estimation par domaine si la contribution de ces unités représente environ 5% de l'estimation totale. Le graphe suivant résume la situation.

Figure1 : Plan d'échantillonnage pour la composante Manufacture



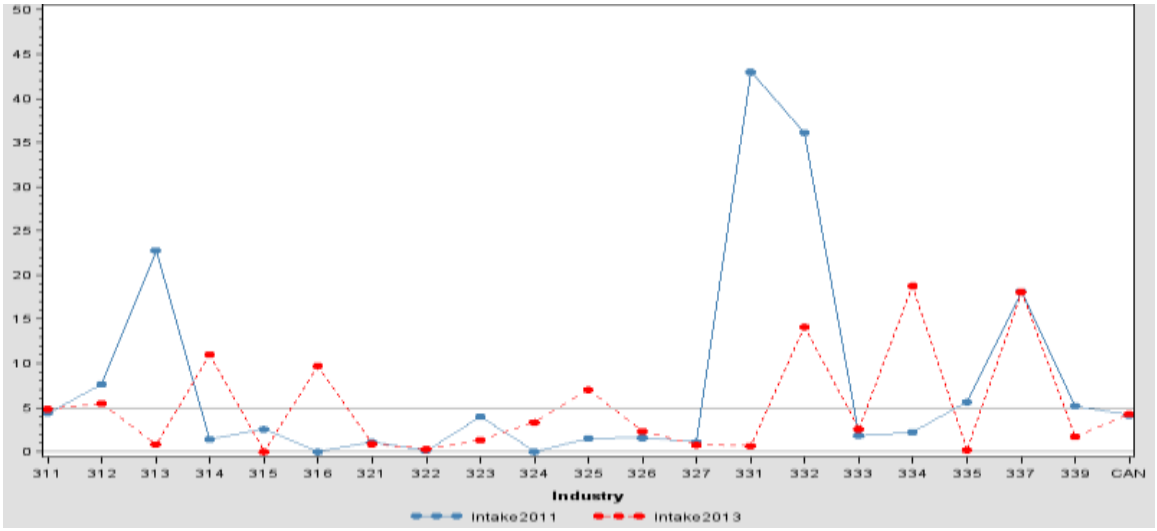
Echantillonnage sur 95% de la population cible (basé sur le revenu)

+

Ajustement par un facteur pour compenser pour les 5%

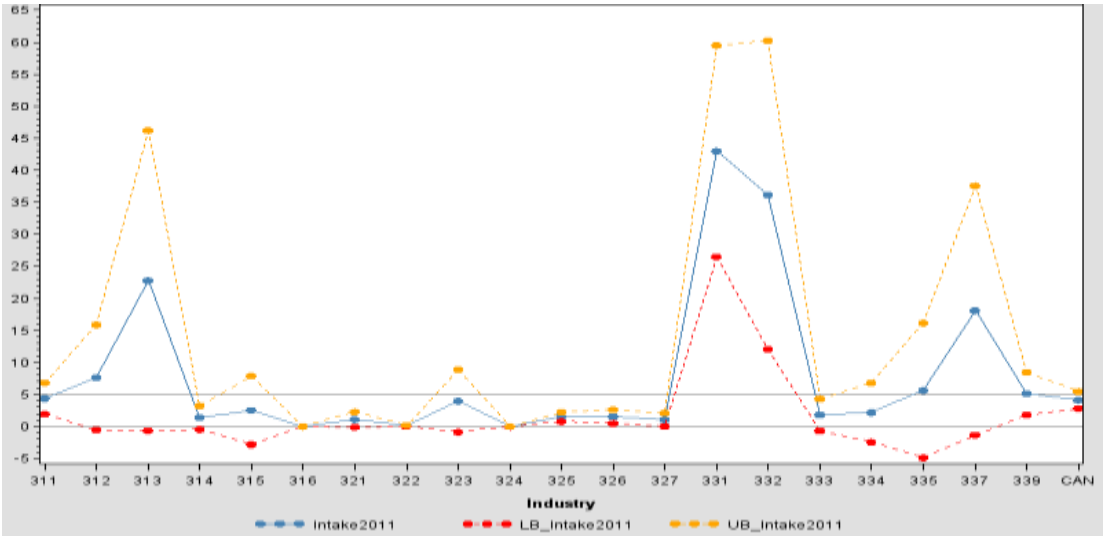
Le graphe Figure2 ci-dessous présente, la contribution des unités additionnelles de 5% pour la variable quantité d'eau prélevée selon différents groupes d'industries pour les trois années. On peut observer que ces contributions sont loin des 5% espérés pour chaque année et varient d'une année à l'autre. Ceci suggère que compenser pour ces 5% par ajustement d'un facteur basé sur le revenu induira un biais difficile à éliminer. Des conclusions similaires sont obtenues quand on réalise l'analyse sur les composantes thermoélectrique et minières.

Figure2 : Contribution (en %) des 5% additionnel pour la variable quantité d'eau prélevée pour les années 2011-2013.



Nous avons produit les intervalles de confiance de l'estimation de la contribution des 5% additionnels. Le graphe Figure3 présente les estimations pour l'année de référence 2011. On peut noter que pour certaines industries, l'intervalle de confiance contient la valeur espérée 5% mais ceci n'est pas le cas pour la plupart. Ce qui suggère un rejet de l'hypothèse que la contribution est égale à 5% pour de nombreuses industries.

Figure3 : Contribution (en %) des 5% additionnelle pour la variable quantité d'eau prélevée pour l'année 2011 avec intervalle de confiance.



2.2.2.3 Cas particulier des centrales thermoélectriques

Nous avons montré dans la section précédente, qu'il est difficile d'estimer sans biais le total incluant la strate à tirage nul. Étant donné que la composante thermoélectrique est un recensement et que nous possédons un long historique des données, nous avons substitué l'ajustement à l'aide de facteur par l'imputation historique par unité. La strate à tirage nul a été tour à tour créée à l'aide du revenu puis d'une variable auxiliaire dérivée de la variable historique quantité d'eau prélevée. Les sections suivantes donnent plus de détails.

2.2.2.3.1 Utilisation du revenu comme variable auxiliaire

Comme dans le scénario précédent, on exclut les unités par l'algorithme de Royce-Maranda avec un seuil de 10%. On procède à un recensement des 90% non exclus. Les unités en-dessous du seuil sont réintégrées dans le processus et sont par la suite traitées comme de la non-réponse et imputées de manière historique. Le Tableau 2 suivant compare les estimations obtenues aux estimations publiées officiellement.

Tableau2: Publication Vs Estimation quantité d'eau avec imputation (2011 & 2013)

	Année 2013		Année 2011	
Région	**Quantité d'eau publiée**	**Nouvel Estimé**	**Quantité d'eau publiée**	**Nouvel Estimé**
Atlantique	2567.53	2557.15	1656.41	1657.2
C Britannique	34.95	34.95	20.31	20.86
Ontario	20844.27	21203.37	19343.96	24193.11
Prairies	2187.58	2678.57	1842.35	2103.32
Québec	0.91	0.91	633.71	633.71
Territoires	0	0	0.48	0

Dans le Tableau 2, même si les estimations sont globalement proches, on peut noter comme dans le cas des prairies pour l'année 2013 par exemple, un grand écart entre la publication et la nouvelle estimation. Une investigation a révélé qu'il s'agit de certaines unités qui fluctuent d'une année à l'autre. Il est plutôt fréquent dans cette enquête que les valeurs de certaines unités ne soient pas stables dans le temps. La nouvelle estimation dans les prairies est nettement supérieure à la publication du fait que certaines unités à revenu faible ont connu une diminution de leur quantité d'eau prélevée. Ces unités ont été classifiées dans la strate à tirage nul. Conséquence, l'imputation historique a entraîné une augmentation des estimations. Il apparaît donc que, compenser la partie non échantillonnée par imputation peut aussi induire un biais dans les estimations.

2.2.2.3.2 Utilisation de la variable historique quantité d'eau en remplacement du revenu

Dans cette section, nous utilisons les valeurs historiques de la variable quantité d'eau comme variable auxiliaire pour définir le seuil d'exclusion dans l'algorithme de Royce-Maranda. Cette stratégie est similaire à celle décrite en 2.2.2.3.1 à l'exception du revenu qui est remplacé par une variable historique dans l'algorithme d'exclusion. Le but est de prendre avantage du fait que la quantité d'eau prélevée au cours d'une année est mieux corrélée à celle de l'année précédente comparée au revenu. Cette stratégie va permettre de réduire le biais au moins pour la variable d'intérêt quantité d'eau prélevé et toutes celles qui lui sont corrélées comme par exemple la quantité d'eau produite. Le Tableau 3 ci-dessous présente un extrait des résultats obtenus pour la région Atlantique. On note encore un écart entre la nouvelle estimation et la publication. Dans ce cas, la valeur d'une unité de cette région a connu une augmentation de plus de 700 millions de mètres cubes entre 2011 et 2013.

Tableau3: Publication Vs Estimation quantité d'eau (2013)

Quantité d'eau utilisée		
Region	**Publiée**	**Nouvelle estimée**
Atlantique	2 568	1 784

Toutes les stratégies explorées ci-dessus sont susceptibles d'induire un biais. Nous sommes présentement en concertation avec l'équipe des spécialistes du sujet du projet afin de définir une stratégie à adopter dans le PISE.

3 Conclusion

Dans cet article, nous avons exploré quelques défis à relever en vue de l'intégration des enquêtes environnementales dans le programme intégré de la statistique des entreprises de Statistique Canada. Toutes les stratégies évaluées ont révélé un problème potentiel de biais quand on applique un plan de sondage avec un seuil d'exclusion. Ce biais potentiel est dû, entre autres au fait que la variable auxiliaire utilisée dans l'algorithme d'exclusion n'est pas bien corrélée avec les variables d'intérêt dans le cas de l'enquête sur l'eau dans les industries.

4 Remerciements

Nous remercions Christopher Tremblay, Kenneth Chu, Nathalie Hamel, Susie Fortier et Wesley Young pour leur effort dans la révision de cet article. Nous remercions également les réviseurs du Colloque pour leurs remarques ayant permis d'améliorer l'article.

Bibliographie

[1] Bee, R. et al. (2007), A Framework for Cut-off Sampling, Journal of Official Statistics, Vol.26, No.4, 2010. pp. 651–671.

[2] Haziza, D. et al (2010), Sampling and estimation in the presence of cut-off sampling, Australian & New Zealand Journal of Statistics, 52: 303–319.

[3] Lahaie, M. et al (2015), Rôles et responsabilités des comités de l'architecture opérationnelle du bureau (CEGAOB et CGAOB): mandat, *Document interne Statistique Canada*.

[4] Royce, D. et Maranda, F. (1998), Groupe de travail sur l'acquisition des données auprès des entreprises, *Rapport interne Statistique Canada*

[5] Yukman, C. et al. (2014), Introduction au programme intégré de la statistique des entreprises (PISE), Note de cours interne à Statistique Canada version 1.1.