
UNE MISE À JOUR SUR L'UTILISATION D'UNE MÉTHODOLOGIE DE COLLECTE DYNAMIQUE POUR LES ENQUÊTES AUPRÈS DES ENTREPRISES

Pierre DAOUST, Jessica ANDREWS, Matei MIREUTA, Loïc MUHIRWA (*)

(*) Statistique Canada, Direction de la méthodologie, Division des méthodes de la statistique économique

Pierre.Daoust@canada.ca, Jessica.Andrews@canada.ca, Matei.Mireuta@canada.ca, Loic.Muhirwa@canada.ca

Mots-clés: collecte active, contrôle sélectif, valeurs prédites, variance totale

Résumé

Le Programme intégré de la statistique des entreprises (Pise) de Statistique Canada (Statistique Canada 2015, [18]), qui sert de cadre normalisé aux enquêtes économiques menées à Statistique Canada, utilise depuis quelques années une méthodologie dynamique pour la collecte et l'analyse des données (Turmelle et coll. 2014, [17]). Cette méthodologie repose sur un système automatisé d'estimations en continu qui permet, à plusieurs reprises lors de la période de production, d'exécuter l'ensemble des processus post-collecte de l'enquête. Ceci permet, entre autres, de comparer certaines mesures de qualité aux cibles correspondantes pour quelques estimations clés de l'enquête, et de recalculer les priorités assignées aux unités de l'échantillon dans le cadre d'une gestion efficace des ressources disponibles pour l'application d'un traitement particulier. Ces priorités sont calculées afin de maximiser l'amélioration espérée des mesures de qualité considérées si le traitement est appliqué. Dans le contexte du Pise, les traitements correspondent aux suivis lors de la collecte des données, nécessaires pour résoudre les cas de non-réponses et d'échecs de contrôles, et aux suivis manuels lors de l'analyse des données. Cette méthodologie est sous-jacente à la composante qui est nommée Indicateurs de qualité et mesures d'impact (IQMI) du système d'estimations en continu.

Le Pise a introduit graduellement l'IQMI lors de la production de ses enquêtes à partir de 2016, et cette composante fait maintenant partie intégrante de la production pour la majorité des enquêtes du Pise, si on exclut les enquêtes avec une très faible taille d'échantillon. Les mesures de qualité initialement utilisées sont le taux de réponse pondéré selon la variable visée par l'estimation et la différence entre les données recueillies et des valeurs prédites préétablies. Pour le moment, un modèle simple qui prend l'hypothèse d'une probabilité à 100% de réussite du traitement est utilisé ; des travaux sont nécessaires afin de développer une estimation automatisée précise et efficace des probabilités de réussite des traitements. Cet article fera le point sur l'utilisation d'IQMI pour les enquêtes du Pise, de l'efficacité de cette approche sur la collecte des données, et discutera d'améliorations prévues. En particulier, ceci inclura un survol d'un projet visant l'utilisation d'une mesure de qualité qui repose sur le concept de variance totale des estimations clés (Beaumont et coll. 2010, [5]). Ceci implique l'automatisation de toutes les étapes nécessaires pour le calcul de cet indicateur de qualité et sa mesure d'impact par extraction de l'information pertinente à partir des composantes d'imputation et d'estimation du système d'estimations en continu.

Abstract

The Statistics Canada Integrated Business Statistics Program (IBSP)(Statistics Canada 2015, [18]), which provides a standardized framework for economic surveys conducted at Statistics Canada, has used for a few years a dynamic methodology for data collection and analysis (Turmelle et al. 2014,

[17]). This methodology is driven by an automated system of rolling estimates that runs, several times during the collection period, the full set of post-collection processes. This allows, amongst other things, a comparison of certain quality measures to corresponding targets for a few key survey estimates, and the recalculation of priorities assigned to sample units within the context of the efficient management of available resources for the application of a specific treatment. These priorities are calculated in order to maximize the expected improvement of the quality measures being considered if the treatment is applied. For IBSP, the treatments correspond to follow-ups at data collection, needed to resolve non-response and failed edits, and to the manual follow-ups conducted at data analysis. This methodology is associated with the quality indicators and measures of impact (QIMI) component of the rolling estimate system.

IBSP has gradually introduced QIMI in the production of its surveys starting in 2016, and this component is now an integral part of production for the majority of IBSP surveys, excluding surveys with a very small sample size. The quality measures initially used are the response rates weighted by the estimate of the targeted variable and the difference between collected data and pre-established predicted values. For now, a simple model that assumes a 100% probability that the treatment will succeed is used; studies are needed to develop an automated process for a precise and efficient estimation of the success probabilities. This paper will provide an update on the use of QIMI for the IBSP surveys, of the impact of this approach on data collection, and will discuss future improvements. In particular, this will include an overview of a project aiming at the use of a quality measure driven by the concept of total variance of key estimates (Beaumont et al. 2010, [5]). This involves the automatization of all steps needed to derive this quality indicator and its measure of impact by extracting the relevant information from the imputation and estimation components of the rolling estimates system.

1. Introduction et contexte

Comme il est décrit dans (Statistique Canada 2015, [18]), en 2009, Statistique Canada a lancé le projet d'examen de l'Architecture opérationnelle du Bureau (AOB). Les principaux objectifs de l'AOB étaient de réaliser des gains d'efficacité, d'améliorer la qualité et d'accroître la réactivité en matière d'exécution des programmes statistiques. Afin d'atteindre les objectifs visés, Statistique Canada a alors entrepris plusieurs projets, dont un grand projet de transformation de ses enquêtes sur la statistique économique, le Programme intégré de la statistique des entreprises (Pise).

Le Pise sert de cadre normalisé aux enquêtes économiques menées à Statistique Canada. Il définit un nouveau modèle et une nouvelle infrastructure d'enquête qui seront utilisés par plus de cent quarante enquêtes entreprises d'ici 2022. Pendant la conception du modèle du Pise, l'équipe s'est concentrée sur six objectifs principaux, que voici :

- améliorer la qualité des données en appliquant des méthodologies et des processus normalisés, en instaurant du contenu harmonisé et en favorisant l'analyse de cohérence;
- alléger le fardeau de réponse;
- moderniser l'infrastructure de traitement des données;
- intégrer la majorité des enquêtes économiques dans le nouveau modèle;
- simplifier et normaliser les processus de manière à réduire les courbes d'apprentissage et à raccourcir les délais;
- réduire les coûts récurrents liés aux aspects opérationnels des enquêtes afin de réaliser des économies.

Pour atteindre ces objectifs, les programmes d'enquête ont dû s'adapter aux exigences particulières du modèle du Pise, mais, en même temps, le modèle devait être conçu de manière suffisamment

souple pour remplir les exigences propres aux différents programmes. Le plus grand défi à relever lors de la mise en œuvre du Pise a été de trouver un équilibre entre ces deux aspects.

Les caractéristiques communes des enquêtes du Pise qui favorisent l'efficacité opérationnelle comprennent :

- l'utilisation pleine et entière du Registre des entreprises comme base de sondage;
- le recours à des questionnaires électroniques comme principal mode de collecte;
- la gestion active de la collecte en fonction des indicateurs de qualité;
- la mise en œuvre d'une stratégie de vérification commune;
- l'application de métadonnées normalisées à la conception des questionnaires, à l'échantillonnage, à la vérification et l'imputation, de même qu'aux processus de répartition des données et d'estimation;
- la gouvernance partagée dans l'ensemble des divisions opérationnelles et spécialisées, y compris la gestion du changement.

Les enquêtes du Pise font appel à ces caractéristiques communes, mais il existe de nombreuses variantes qui répondent aux exigences propres à chaque enquête. Par exemple, le modèle est conçu de manière suffisamment souple pour traiter des enquêtes de différentes fréquences, notamment les enquêtes mensuelles, trimestrielles et annuelles et ayant différents modèles de couverture, comme des enquêtes à l'échelle de l'économie, des enquêtes axées sur l'industrie et des enquêtes fondées sur les activités.

Une description détaillée de ces caractéristiques communes est disponible dans (Statistique Canada 2015, [18]). Certaines ont été décrites et présentées au cours des années qui ont précédé ou depuis le lancement du Pise; consulter par exemple (Godbout 2011, [9]), (Godbout et coll. 2011, [10]), (Turmelle et coll. 2012, [16]), (Mills et coll. 2013, [13]), (Turmelle et Beaucage 2013, [15]), (Turmelle et coll. 2014, [17]), (Andrews et coll. 2016, [1]), (Mireuta et coll. 2017, [12]), ou (Andrews et coll. 2018, [2]) pour une description du modèle de traitement des données du Pise et de sa stratégie de gestion active de la collecte et de l'analyse. Le présent article se concentrera sur la stratégie de gestion active de la collecte et de l'analyse. Un survol de la méthodologie associée à cette stratégie sera donné aux sections 2 et 3 et de sa mise en œuvre à la section 4, inspirée fortement du guide méthodologique des mesures de qualité du Pise (Bosa et Godbout 2014, [7]). La section 5 discutera de l'impact de l'utilisation de cette stratégie pour les enquêtes du Pise. Finalement les plans futurs seront discutés à la section 6.

2. Gestion active de la collecte et de l'analyse

2.1. Modèle d'estimations en continu du Pise

La gestion active de la collecte et de l'analyse est une expression qui signifie que les efforts de collecte et d'analyse sont ajustés de manière dynamique en fonction des données déjà reçues. La pierre angulaire de cette gestion active dans le Pise est le modèle des estimations en continu (EC), selon lequel on produit des estimations de manière itérative pendant la période de collecte jusqu'à ce qu'on ait atteint un niveau de qualité acceptable. Ce modèle a été décrit originalement dans (Godbout et coll. 2011, [10]).

Chaque itération sous ce modèle implique l'intégration des données de collecte et des données administratives disponibles à ce moment, le suivi du traitement automatisé des données afin de produire les estimations pour l'enquête ainsi que des indicateurs de qualités correspondants. Les indicateurs de qualité (IQ) de variables clés par domaine d'estimation ainsi que les mesures de l'impact (MI) de chaque unité fournissent l'information nécessaire pour déterminer quelles unités

prioriser pour les suivis de collecte et pour l'analyse ou la vérification manuelle d'unités dans la stratégie d'analyse du Pise.

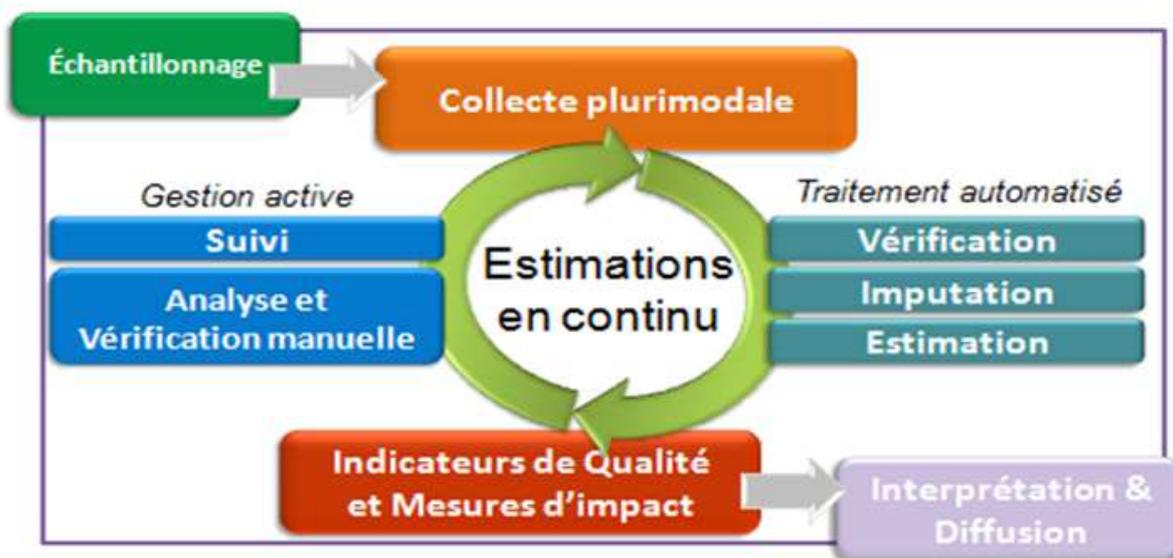


Figure 1 : Modèle d'estimation en continu du Pise

Lorsque tous les objectifs de qualité sont atteints pour une enquête particulière, la collecte active est fermée et le suivi peut prendre fin; entre temps, les ressources chargées du suivi ou de l'analyse et de la vérification manuelle sont affectées aux unités considérées avoir un impact important sur les estimations clés et leur qualité. Grâce au modèle des estimations en continu et à la gestion active de la collecte, on s'attend à une réduction globale des activités de suivi de la collecte, ce qui allégera le fardeau de réponse et améliorera la cohérence de la stratégie afin de cibler la qualité dans les différents domaines d'estimation.

2.2. Estimation et imputation dans le Pise

On suppose que l'enquête désire estimer, à partir d'un échantillon s , les totaux $t_{y,d} = \sum_{k \in U} y_{vdk}$ d'une population U pour différentes variables y_v et plusieurs domaines d'intérêt d , où y_{vdk} représente la valeur associée à l'unité k pour une variable y_v et un domaine d'intérêt d . Pour ce faire, plusieurs étapes sont nécessaires. De brefs aperçus de l'imputation et de l'estimation seront présentés respectivement aux sections 2.2.1 et 2.2.2.

2.2.1. Estimation

L'estimateur du total après allocation, imputation et calage est donné par :

$$\begin{aligned} \hat{t}_{y,d} &= \sum_{k \in s_{yy}} w_k f_{y,dk} y_{vk}^{(R)} + \sum_{k \in s_{my}} w_k f_{y,dk} y_{vk}^{(I)} \\ &= \sum_{k \in s} w_k f_{y,dk} y_{vk}^* \end{aligned} \tag{1}$$

où y_{vk}^* correspond à une valeur rapportée $y_{vk}^{(R)}$ (si $k \in s_{ry}$, l'ensemble des unités répondantes¹) ou une valeur imputée $y_{vk}^{(I)}$ (si $k \in s_{my}$, l'ensemble des unités imputées) de la variable d'intérêt y_v pour l'unité k . Le poids final utilisé à l'estimation w_k tient compte des probabilités d'inclusion dans l'échantillon et des ajustements effectués à l'estimation. La valeur imputée $y_{vk}^{(I)}$ est donnée par $y_{vk}^{(I)} = \hat{\mu}_{y_v, k}$ selon les méthodes et modèles décrits dans les sections 2.2.2 et 2.2.3. Le facteur d'allocation $f_{y_v, dk} \in [0, 1]$ est une constante provenant du Registre des entreprises, disponible pour l'ensemble de la population, attribuant la portion de la valeur y_{vk} pour l'unité k revenant au domaine d pour la variable d'intérêt y_v . On note que $f_{y_v, dk} = 0$ si l'unité k ne contribue pas au domaine d .

2.2.2. Imputation

Ce qui suit décrit brièvement les modèles d'imputation pour lesquels on peut calculer la variance due à l'imputation avec la méthodologie derrière le **Système d'estimation de la variance** due à la non-réponse et à l'imputation (Sevani) qui a été développé et intégré au système généralisé d'estimation G-EST par Statistique Canada (Beaumont et coll. 2010, [5]). Dans le Pise, plusieurs méthodes et modèles d'imputation seront appliqués de façon séquentielle pour une même variable (imputation composite). L'article (Beaumont & Bissonnette 2011, [4]) contient plus d'information et une généralisation pour des méthodes d'imputation composite. Cette généralisation n'est pas nécessaire dans ce document et donc la notation sera légèrement simplifiée en comparaison avec l'article.

Tout comme Sevani, le cadre méthodologique des IQs et des MIs du Pise requiert des méthodes d'imputation linéaires, c.-à-d. qu'on peut écrire une valeur imputée $y_{vk}^{(I)}$ comme une combinaison linéaire des valeurs rapportées par les autres unités:

$$y_{vk}^{(I)} = \varphi_{0, y_v, k} + \sum_{k' \in s_{ry}} \varphi_{y_v, k'k} y_{vk'}^{(R)} \quad (2)$$

Les quantités $\varphi_{0, y_v, k}$ et $\varphi_{y_v, k'k}$ ne dépendent pas des valeurs de la variable d'intérêt y_v , mais peuvent dépendre de s et s_{ry} .

On définit la quantité $W_{y_v, dk}$ qui représente le poids de compensation (ou d'ajustement) pour contrer la non-réponse portée par l'unité $k' \in s_{ry}$ ayant une valeur rapportée $y_{vk'}^{(R)}$ allouée au domaine d pour la variable d'intérêt y_v . $W_{y_v, dk}$ et $W_{0, y_v, d}$ sont définis de la manière suivante :

$$\begin{aligned} W_{y_v, dk'} &= \sum_{k \in s_{my}} w_k f_{y_v, dk} \varphi_{y_v, k'k} \\ W_{0, y_v, d} &= \sum_{k \in s_{my}} w_k f_{y_v, dk} \varphi_{0, y_v, k} \end{aligned} \quad (3)$$

On peut donc réécrire l'équation (1) qui est définie à la section 2.2.1 de la manière suivante

¹Par unités répondantes on indique les unités dans l'échantillon pour lesquelles la valeur est considérée rapportée, peu importe la source (collecte, donnée administrative, imputation manuelle).

$$\begin{aligned}\hat{t}_{y_v d} &= \sum_{k \in s_{ryv}} w_k f_{y_v dk} y_{vk}^{(R)} + \sum_{k \in s_{myv}} w_k f_{y_v dk} y_{vk}^{(I)} \\ &= W_{0y_v d} + \sum_{k \in s_{ryv}} y_{vk}^{(R)} (w_k f_{y_v dk} + W_{y_v dk})\end{aligned}$$

2.2.3. Modèle d'imputation

On considère le modèle d'imputation suivant qui décrit la relation entre la variable y_v et le vecteur de variables auxiliaires observées \mathbf{x}^{obs} :

$$\begin{aligned}E_m(y_{vk} | \mathbf{X}^{obs}) &= \mu_{y_v k} \\ V_m(y_{vk} | \mathbf{X}^{obs}) &= \sigma_{y_v k}^2 \\ cov_m(y_{vk}, y_{vk'} | \mathbf{X}^{obs}) &= 0,\end{aligned}\tag{4}$$

où $k \neq k'$ et $k, k' \in U$. La matrice \mathbf{X}^{obs} contient tous les vecteurs observés \mathbf{x}^{obs} . Le m dans l'équation (4) signifie que l'espérance, la variance et la covariance sont par rapport au modèle d'imputation.

Les quantités $\mu_{y_v k}$ et $\sigma_{y_v k}^2$ seront estimées respectivement par $\hat{\mu}_{y_v k}$ et $\hat{\sigma}_{y_v k}^2$. On suppose que ces estimateurs sont sans biais et convergents par rapport au modèle m .

Les quantités définies par le modèle d'imputation seront utiles à la section 3.4 quand les IQs et les MIs seront définis.

3. Description des indicateurs de qualité et mesures d'impact

3.1. Concepts préliminaires

La gestion active se fait d'abord en définissant 3 concepts de base : la liste des estimations clés à gérer, les facteurs d'importance et les cibles de qualité à atteindre. Ces quantités sont définies avant la première itération du processus d'estimation en continu (EC) et demeurent statiques, en principe, tout au long d'un cycle de l'enquête. L'objectif de la gestion active est d'amener chacune des estimations clés à atteindre les cibles de qualité, et ce, de la façon la plus efficace possible.

3.1.1. Estimations clés

En premier lieu, un ensemble de variables clés \mathbf{K}_V est identifié. \mathbf{K}_V est un sous-ensemble des variables y_v ($v = 1$ à V) de l'enquête. Ensuite, un ensemble de domaines clés \mathbf{K}_D est identifié. \mathbf{K}_D est un sous-ensemble des domaines d'intérêt de l'enquête. Finalement, l'ensemble $\mathbf{K}_E = \mathbf{K}_V \times \mathbf{K}_D$ d'estimations clés est simplement le produit cartésien entre l'ensemble \mathbf{K}_V des variables clés et celui \mathbf{K}_D des domaines clés, bien qu'en pratique un sous-ensemble de ce produit cartésien peut être utilisé.

La liste des estimations clés est fournie par les agents spécialisés en suivant certaines lignes directrices. En particulier,

Variables clés :

- Elles devraient faire l'objet de vérifications obligatoires pour faciliter le suivi des cas rejetés au contrôle.
- Les variables remplacées par des données fiscales, et qui lorsqu'imputées sont considérées comme répondues, ne sont pas éligibles au suivi effectué au cours de la collecte. Cependant, elles pourraient être utilisées aux fins de l'analyse.
- Les variables doivent figurer dans les fichiers d'estimation aux niveaux macro et micro.
- Elles ne doivent pas prendre de valeurs négatives.
- Il doit être possible de modéliser la prédiction du cycle courant pour ces variables selon des données historiques ou auxiliaires, au minimum au niveau du total des domaines d'intérêt.
- Les variables rarement déclarées (populations rares) peuvent causer des problèmes; les unités qui déclarent ces variables se verront fréquemment attribuer une plus grande priorité.

Domaines clés :

- On recommande d'éviter les domaines dont le niveau est plus détaillé que les cellules d'échantillonnage. La taille de l'échantillon n'étant pas contrôlée; il se pourrait qu'aucune unité n'ait été sélectionnée.
- Les petits domaines doivent être évités, sauf s'ils sont très importants, puisque comme pour les populations rares les unités de ces domaines se verront fréquemment attribuer une plus grande priorité.

Les estimations clés devraient faire partie des estimations diffusées ou publiées. Il y a une perte d'efficacité s'il y a trop de variables clés.

3.1.2. Facteurs d'importance

Un facteur d'importance (FI) est un poids associé à une estimation clé. Le FI permet de quantifier l'importance relative des estimations clés les unes par rapport aux autres.

Dans le contexte du Pise, on définit le facteur d'importance de la manière suivante :

$$FI_{y_v, d} = FI_{y_v, d_1 \dots d_{\Xi}} = \left(\frac{\alpha_v t_{x_v}}{t} \right)^{p_0} \left(\frac{t_{x_v, d_1}}{t_{x_v}} \right)^{p_1} \dots \left(\frac{t_{x_v, d_1 d_2 \dots d_{\xi}}}{t_{x_v, d_1 d_2 \dots d_{\xi-1}}} \right)^{p_{\xi}} \dots \left(\frac{t_{x_v, d}}{t_{x_v, d_1 d_2 \dots d_{\Xi-1}}} \right)^{p_{\Xi}} \quad (5)$$

Dans l'équation (5), α_v représente un paramètre d'ajustement ou un changement d'échelle pour être en mesure de comparer adéquatement les diverses variables y_v , $d^{(\xi)} = d_{\xi}$ fait référence à un domaine défini selon la valeur d_{ξ} d'une classification $d^{(\xi)}$, t_{x_v} représente le total pour la variable auxiliaire x_v qui est une variable corrélée avec y_v et $t = \sum_v \alpha_v t_{x_v}$. Les valeurs p_0 , p_1 , p_{ξ} et p_{Ξ} sont des nombres réels entre 0 et 1. L'exposant p_0 et le paramètre α_v permettent de définir les importances relatives entre les différentes variables clés y_v . La valeur p_{ξ} permet de définir l'importance des domaines clés $d^{(1)}, d^{(2)}, \dots, d^{(\xi)} = d_1, d_2, \dots, d_{\xi}$ pour toutes les valeurs d_{ξ} que peut prendre la classification $d^{(\xi)}$ étant donné que les domaines $d^{(1)}, d^{(2)}, \dots, d^{(\xi-1)} = d_1, d_2, \dots, d_{\xi-1}$ et la

variable y_v , sont fixés. En d'autres termes, p_ξ permet de définir l'importance des différents domaines $d^{(\xi)} = d_\xi$ à l'intérieur de la sous-population $U_{y, d_1 d_2 \dots d_{\xi-1}}$. Pour mieux comprendre la valeur p_ξ , on peut faire un parallèle avec la répartition de puissance en échantillonnage présentée dans (Bankier 1988, [3]).

- Une valeur de p_ξ près de 0 signifie que l'on veut une importance similaire pour les différents domaines $d^{(\xi)} = d_\xi$ à l'intérieur de la sous-population $U_{y, d_1 d_2 \dots d_{\xi-1}}$.
- Plus la valeur de p_ξ est proche de 1, plus l'importance pour les différents domaines $d^{(\xi)} = d_\xi$ sera proportionnelle à $t_{x, d_1 d_2 \dots d_\xi}$ à l'intérieur de la sous-population $U_{y, d_1 d_2 \dots d_{\xi-1}}$.

Pour les enquêtes du Pise, nous ciblons généralement de 4 à 6 variables clés et deux dimensions, soient la classification géographique $d^{(1)}$ et la classification industrielle $d^{(2)}$ basée sur le code Système de classification des industries de l'Amérique du Nord (SCIAN).

3.1.3. Cibles de qualité

Une cible de qualité (CQ) représente la valeur de l'indicateur de qualité qui est visé pour une estimation clé et une mesure de qualité donnée. La qualité visée est directement liée au facteur d'importance; plus le facteur d'importance est élevé, plus la qualité visée sera meilleure.

Pour la deuxième itération d'une enquête dans le Pise, des estimations et indicateurs de qualité historiques sont disponibles et peuvent servir de base pour déterminer les cibles de qualité pour les prochaines itérations. Toutefois, pour le premier cycle dans le Pise, des cibles de qualité généralement conservatrices sont établies à partir de l'information historique sous le plan d'enquête précédent.

3.2. Qualité - Définition

Les définitions générales et conceptuelles sont présentées dans cette section, mais les expressions mathématiques pour quelques mesures de qualité seront présentées à la section 3.4.

3.2.1. Mesure de qualité (MQ)

Comme son nom l'indique, ceci représente une mesure pour évaluer la qualité des estimations. Deux types de MQ sont définis ci-dessous.

1. MQ décroissante : la qualité est meilleure quand la valeur de la MQ (c.-à-d. l'indicateur de qualité) est petite. Par exemple, les divers taux d'imputation et le coefficient de variation (CV) sont des MQ décroissantes.
2. MQ croissante : la qualité est meilleure quand la valeur de la MQ est grande. On suppose que la valeur maximale est de 1 et donc qu'un IQ de 1 est la meilleure qualité possible. Par exemple, les taux de réponse sont des MQ croissantes dont la valeur maximale est de 1.

Dans ce document, sauf lorsque noté, les concepts principaux sont définis selon des MQs décroissantes. Il suffit d'appliquer des changements de variables pour appliquer cette théorie aux MQs croissantes.

3.2.2. Indicateur de qualité (IQ)

3.2.2.1. IQ local

L'indicateur de qualité ($IQ(\hat{\theta}_{y,d})$) local représente une valeur numérique associée à une mesure de qualité pour l'estimation $\hat{\theta}_{y,d}$ du paramètre $\theta_{y,d}$ (par exemple le total de la variable y_v pour le domaine d). La valeur $IQ(\hat{\theta}_{y,d})$ quantifie la qualité pour une mesure de qualité donnée et une estimation clé (variable et domaine) donnée. Souvent, on utilise une estimation $I\hat{Q}(\hat{\theta}_{y,d})$ car $IQ(\hat{\theta}_{y,d})$ n'est pas disponible. Par exemple, on ne connaît pas la variance, et donc le coefficient de variation, d'un estimateur du total mais plutôt une estimation calculée à l'aide d'un échantillon. Dans le cas où il est possible de calculer $IQ(\hat{\theta}_{y,d})$ comme pour les taux de réponse et d'imputation non pondérés, nous allons supposer que $I\hat{Q}(\hat{\theta}_{y,d}) = IQ(\hat{\theta}_{y,d})$ afin de simplifier la notation.

3.2.2.2. IQ global

Afin de simplifier l'analyse de l'ensemble des indicateurs de qualité pour toutes les estimations clés $\hat{\theta}_{y,d}$, il est intéressant de les combiner en une statistique globale. La fonction de distance retenue est une moyenne quadratique pondérée correspondant à la fonction objective de puissance utilisée à l'échantillonnage lors de la répartition multivariée de l'échantillon. Pour plus de détails sur la fonction objective utilisée à l'échantillonnage, voir (Turmelle et coll. 2012, [16]). Ainsi, l'estimation de l'indicateur de qualité global $I\hat{Q}^G(\hat{\theta})$, où $\hat{\theta}$ représente le vecteur de toutes les estimations clés, associé à une MQ donnée est définie par :

$$I\hat{Q}^G(\hat{\theta}) = \sqrt{\frac{\sum_{(v,d) \in K_E} (FI_{y,d} I\hat{Q}(\hat{\theta}_{y,d}))^2}{\sum_{(v,d) \in K_E} (FI_{y,d})^2}} \quad (6)$$

où $FI_{y,d}$ est le facteur d'importance associé à la variable y_v et au domaine d . $I\hat{Q}^G(\hat{\theta})$ est un indicateur unique au niveau de l'enquête pour une mesure de qualité donnée.

3.2.3. Distance de qualité (DQ)

3.2.3.1. DQ locale

Avant de relier les indicateurs de qualité et la méthodologie de la gestion active, on définit la mesure de distance de qualité locale $D\hat{Q}(\hat{\theta}_{y,d})$ associée à une MQ donnée pour l'estimation clé $\hat{\theta}_{y,d}$.

$$D\hat{Q}(\hat{\theta}_{y,d}) = \begin{cases} \max\left(\frac{I\hat{Q}(\hat{\theta}_{y,d}) - CQ(\hat{\theta}_{y,d})}{I\hat{Q}(\hat{\theta}_{y,d})}, 0\right) & \text{si } I\hat{Q}(\hat{\theta}_{y,d}) > 0 \\ 0 & \text{si } I\hat{Q}(\hat{\theta}_{y,d}) = 0 \end{cases} \quad (7)$$

$CQ(\hat{\theta}_{y,d})$ représente la cible de qualité associée à $\hat{\theta}_{y,d}$ pour une mesure de qualité donnée, soit la

valeur de $I\hat{Q}(\hat{\theta}_{y,d})$ que l'on veut atteindre avant la fin de l'enquête.

3.2.3.2. DQ globale

Les distances locales $D\hat{Q}_{y,d}$ définies par l'équation (7) peuvent être combinées en une distance de qualité globale $D\hat{Q}^G$ de la façon suivante :

$$D\hat{Q}^G(\hat{\theta}) = \sqrt{\frac{\sum_{(v,d) \in K_E} (FI_{y,d} D\hat{Q}(\hat{\theta}_{y,d}))^2}{\sum_{(v,d) \in K_E} FI_{y,d}^2}} \quad (8)$$

On peut remarquer que la distance globale de qualité $D\hat{Q}^G(\hat{\theta})$ est positive tant qu'il y a des estimations clés pour lesquelles les cibles de qualité ne sont pas atteintes et diminue lorsque les indicateurs s'approchent de leurs cibles. Cette distance sera nulle dès que chaque cible de qualité est atteinte, et constitue donc un indicateur essentiel pour la gestion active de la collecte.

3.3. Mesure d'impact (MI)

Certains concepts doivent être introduits avant de définir ce qu'est une MI.

3.3.1. Impact sur la statistique $\hat{\xi}_{y,d}$ du traitement T sur une unité k

La définition générale de l'impact $\delta_{Tk}(\hat{\xi}_{y,d})$ d'un traitement T sur une unité k pour une statistique $\hat{\xi}_{y,d}$ conditionnellement à s et s_{ry_v} est donnée par :

$$\delta_{Tk}(\hat{\xi}_{y,d}) = \hat{\xi}_{y,d} - E_{Tk}[\hat{\xi}_{y,d} | s, s_{ry_v}] \quad (9)$$

où $\hat{\xi}_{y,d}$ représente l'estimation de la statistique $\xi_{y,d}$ pour le domaine d et une variable y_v . La nouvelle valeur de $\hat{\xi}_{y,d}$ si le traitement T sur l'unité k est un succès est désignée par $\hat{\xi}_{y,d}^{(k)}$, que l'on suppose indépendante du traitement T . Ainsi, $E_{Tk}[\hat{\xi}_{y,d} | s, s_{ry_v}]$ peut être exprimée à l'aide de l'expression ci-dessous :

$$E_{Tk}[\hat{\xi}_{y,d} | s, s_{ry_v}] = \hat{\xi}_{y,d}^{(k)} P_{Tk} + \hat{\xi}_{y,d} (1 - P_{Tk})$$

où P_{Tk} est la probabilité que le traitement T sur l'unité k soit un succès. Par exemple, si le traitement T est le suivi téléphonique pour la non-réponse alors P_{Tk} représente la probabilité de convertir une unité k non répondante en une unité répondante. On peut donc réécrire l'équation (9) de la manière suivante :

$$\delta_{Tk}(\hat{\xi}_{y,d}) = P_{Tk} (\hat{\xi}_{y,d} - \hat{\xi}_{y,d}^{(k)}) \quad (10)$$

Pour les premières années de mise en œuvre du Pise, il n'y aura pas d'outil de modélisation de la propension de réponse afin d'estimer P_{Tk} , forçant à supposer que tous les traitements seront un succès, c'est-à-dire que $P_{Tk} = 1$. Ainsi, on utilisera l'équation suivante :

$$\delta_{Tk}(\hat{\xi}_{y,d}) = \hat{\xi}_{y,d} - \hat{\xi}_{y,d}^{(k)} \quad (11)$$

3.3.2. Mesure d'impact locale

Une mesure d'impact locale $MI_{Tk}(\hat{\theta}_{y,d})$ est toujours associée à une mesure de qualité et représente l'amélioration potentielle de l'indicateur de qualité $I\hat{Q}(\hat{\theta}_{y,d})$ conditionnellement au traitement T sur une unité k . La mesure d'impact sous le traitement T pour l'unité k sera donnée par

$$MI_{Tk}(\hat{\theta}_{y,d}) = \delta_{Tk}(I\hat{Q}(\hat{\theta}_{y,d})) = P_{Tk}(I\hat{Q}(\hat{\theta}_{y,d}) - I\hat{Q}(\hat{\theta}_{y,d})^{(k)}) \quad (12)$$

3.3.3. Mesure d'impact globale

Une mesure d'impact globale $MI_{Tk}^G(\hat{\theta})$ sur une estimation $\hat{\theta}$ pour un traitement T sur une unité k est obtenue en combinant les différentes MIs locales $MI_{Tk}(\hat{\theta}_{y,d})$. La référence (Hedlin 2008, [11]) présente diverses méthodes pour combiner des scores locaux. La mesure d'impact globale représente l'importance qu'a une unité sur l'ensemble des estimations d'une enquête pour une mesure de qualité donnée.

La mesure d'impact globale est définie à partir des facteurs d'importance $FI_{y,d}$, des scores $MI_{Tk}(\hat{\theta}_{y,d})$ locaux (voir équation (12)) et des distances locales $D\hat{Q}(\hat{\theta}_{y,d})$ (voir équation (7)). Cette mesure d'impact globale est donnée par l'équation suivante :

$$MI_{Tk}^G(\hat{\theta}) = \sqrt{\frac{\sum_{(v,d) \in K_E} (FI_{y,d} D\hat{Q}(\hat{\theta}_{y,d}) MI_{Tk}(\hat{\theta}_{y,d}))^2}{\sum_{(v,d) \in K_E} FI_{y,d}^2}} \quad (13)$$

Dans cette mesure, les facteurs d'importance et les distances locales donnent plus d'importance aux MIs locales dont l'estimation clé est considérée importante ou dont l'IQ locale est loin de la cible.

3.3.4. Listes de priorisation

Une fois les MIs globales calculées, nous sommes en mesure de créer la liste des unités assignées et priorisées pour le traitement T . Une liste contenant les unités priorisées peut être créée en utilisant plusieurs mesures de qualité et donc en combinant plusieurs MIs globales si nécessaire.

Pour le Pise, une combinaison linéaire des MI globales est utilisée, ainsi le score final est calculé selon

$$MI_{Tk}^F(\hat{\theta}) = \sum_{MQ} \alpha_{MQ} MI_{Tk, MQ}^G(\hat{\theta}) \quad (14)$$

Pour le moment les valeurs de α_{MQ} à utiliser sont dérivées empiriquement selon les mesures de qualité combinées.

Au départ, il faut noter que certaines unités ne sont pas admissibles au traitement T , par exemple pour des raisons logistiques. Ces unités ne sont pas prises en compte dans la priorisation pour le traitement T .

Finalement, les unités qui représentent les 5% des scores les plus élevés sont classées selon le groupe de priorité 1, le prochain 10% représente le groupe de priorité 2, et ainsi de suite. Les efforts de collecte sont gérés selon ces groupes de priorité. En général les 3 premiers groupes de priorité sont

visés pour les suivis reliés à la non-réponse totale des enquêtes, alors que les 2 premiers groupes de priorité sont visés pour les suivis reliés à l'échec de contrôles de collecte.

3.4. Mesures de qualité

Dans cette section, les IQs et les MIs locaux seront définis pour différentes mesures de qualité lorsque le total est le type d'estimation considéré ($\hat{\theta}_{y,d} = \hat{t}_{y,d}$). On suppose aussi que la probabilité P_{Tk} que le traitement T soit un succès sur l'unité k est de 1. Pour généraliser ce qui suit, les MIs peuvent être multipliés par P_{Tk} dans le cas où cette probabilité peut être différente de 1.

3.4.1. Taux de réponse

Le taux de réponse est une mesure de qualité croissante. L'IQ dans un domaine d pour la variable y_v est donné par la formule suivante :

$$IQ(\hat{t}_{y,d}) = \frac{\sum_{k \in S_{y_v}} f_{y,d,k} w_k y_{vk}^{(R)}}{\sum_{k \in S} f_{y,d,k} w_k y_{vk}^*} = \frac{\sum_{k \in S_{y_v}} f_{y,d,k} w_k y_{vk}^{(R)}}{\hat{t}_{y,d}} \quad (15)$$

Cet indicateur de qualité représente la proportion de l'estimation provenant des unités répondantes. Selon l'équation (12) la MI devient :

$$\hat{MI}_k(\hat{t}_{y,d}) = \frac{f_{y,d,k} (1 - I_{ry,k}) w_k y_{vk}^{(I)}}{\sum_{k \in S} f_{y,d,k} w_k y_{vk}^*} = \frac{f_{y,d,k} (1 - I_{ry,k}) w_k y_{vk}^{(I)}}{\hat{t}_{y,d}} \quad (16)$$

où $I_{ry,k}$ est un indicateur de réponse d'item de la variable y_v pour l'unité k . Ceci signifie qu'une unité non répondante k pour la variable y_v et contribuant au domaine d augmentera le taux de réponse d'environ $f_{y,d,k} w_k y_{vk}^{(I)} / \hat{t}_{y,d}$ si elle est convertie en une unité répondante et si la valeur imputée $y_{vk}^{(I)}$ n'est pas trop loin de la valeur rapportée $y_{vk}^{(R)}$. En d'autres termes, la MI des unités non répondantes contribuant à une estimation clé donnée sera proportionnelle à $w_k y_{vk}^{(I)}$.

Le cas général de cette mesure est aussi appelé Taux de réponse pondéré de la variable clé. En pratique, une mesure robuste de cet indicateur de qualité est utilisée, alors que des valeurs répondues extrêmes sont remplacées par des valeurs prédites.

Voici 2 cas particuliers

- Pour le taux de réponse non pondéré,
 - on pose $w_k = y_{vk}^{(R)} = y_{vk}^{(I)} = 1$;
 - L'IQ représente la proportion des unités répondantes dans l'échantillon;
 - La MI pour une estimation clé sera la même pour les unités non répondantes contribuant à cette estimation clé.
- Taux de réponse pondéré,

- on pose $y_{vk}^{(R)} = y_{vk}^{(I)} = 1$ et on utilise le poids w_k ;
- L'IQ représente la proportion pondérée des unités répondantes dans l'échantillon;
- La MI des unités non répondantes contribuant à une estimation clé donnée sera proportionnelle au poids w_k .

Des formulations similaires selon le taux d'imputation, une mesure de qualité décroissante, sont également disponibles.

3.4.2.Écart relatif par rapport aux valeurs prédites (ERVP)

L'ERVP est une mesure de qualité décroissante. L'erreur prédite relative mesure la déviation entre l'estimation et la prédiction. Le but de cette mesure de qualité est de mesurer l'écart entre les données utilisées pour produire les estimations et les données attendues. Les données attendues sont des prédictions faites pour les variables clés sur l'ensemble des unités échantillonnées.

Les valeurs prédites sont produites à partir de données historiques, de données administratives et autres modèles de prédiction avant même le début de la période de collecte. L'IQ dans un domaine d pour la variable y_v est donné par la formule suivante :

$$I\hat{Q}(\hat{t}_{y,d}) = \frac{\sum_{k \in s} |f_{y,dk} w_k (y_{vk}^* - \tilde{y}_{vk}^{(P)})|}{\sum_{k \in s} f_{y,dk} w_k y_{vk}^*} = \frac{\sum_{k \in s} |f_{y,dk} w_k (y_{vk}^* - \tilde{y}_{vk}^{(P)})|}{\hat{t}_{y,d}} \quad (17)$$

où $\tilde{y}_{vk}^{(P)}$ représente la valeur prédite de l'unité k pour la variable y_v .

On définit selon l'équation (12) la MI de la manière suivante :

$$\hat{M}I_k(\hat{t}_{y,d}) = \frac{|(f_{y,dk} w_k (y_{vk}^* - \tilde{y}_{vk}^{(P)}))|}{\hat{t}_{y,d}} \quad (18)$$

Dans le cas présent l'IQ deviendra $\left(\sum_{l \neq k \in s} |f_{y,dl} w_l (y_{vl}^* - \tilde{y}_{vl}^{(P)})| \right) / \hat{t}_{y,d}$ si l'unité k est vérifiée et que l'on enlève son impact sur cet IQ.

Les MIs selon ERPV sont automatiquement mises à 0 dans le contexte de la gestion active de la collecte quand un agent spécialisé vérifie un enregistrement. L'idée derrière ceci est qu'on ne veut pas avoir un score élevé à cause de l'erreur prédite pour une valeur qui a été vérifiée. Ceci n'est pas fait pour la gestion active de l'analyse, bien que l'agent spécialisé ait également accès à la version utilisée pour la gestion de la collecte.

3.4.3.Coefficient de variation (CV) basée sur la variance totale²

L'indicateur de qualité et la mesure d'impact seront présentés dans un premier temps comme pour les trois autres mesures de qualité précédentes. Toutefois, la notation et les équations seront expliquées ultérieurement dans cette section en raison de la complexité de cette MQ.

²Par variance totale on réfère à la variabilité due à l'échantillonnage et à l'imputation

Le coefficient de variation permet de quantifier la variabilité de l'estimateur. L'IQ dans un domaine d pour la variable y_v est donné par la formule suivante :

$$IQ(\hat{t}_{y,d}) = \frac{\sqrt{\hat{V}_{TOT}(\hat{t}_{y,d})}}{\hat{t}_{y,d}} \quad (19)$$

où $\hat{V}_{TOT}(\hat{t}_{y,d})$ sera défini plus tard par l'équation (22).

La MI est donnée par l'expression suivante :

$$\hat{MI}_k(\hat{t}_{y,d}) = \frac{\sqrt{\hat{V}_{TOT}(\hat{t}_{y,d})}}{\hat{t}_{y,d}} - \frac{\sqrt{\hat{V}_{TOT}(\hat{t}_{y,d}) - \delta_k(\hat{V}_{TOT}(\hat{t}_{y,d}))}}{\hat{t}_{y,d}} \quad (20)$$

où

$$\delta_k(\hat{V}_{TOT}(\hat{t}_{y,d})) = \delta_k(\hat{V}_{ORD}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{DIF}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{NR}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{MIX}(\hat{t}_{y,d}))$$

Cette mesure de qualité est plus complexe que les mesures précédentes. Une brève introduction sur cette théorie sera donnée avant d'obtenir $\delta_k(\hat{V}_{TOT}(\hat{t}_{y,d}))$. Les termes de l'expression (20) seront définis ultérieurement dans la prochaine section.

3.4.3.1. Introduction sur la variance totale

L'erreur quadratique moyenne totale se décompose en 4 termes selon l'équation suivante :

$$E_{TOT}(\hat{t}_{y,d} - t_{y,d})^2 = \hat{V}_{SAM}(\hat{t}_{y,d}) + \hat{V}_{NR}(\hat{t}_{y,d}) + \hat{V}_{MIX}(\hat{t}_{y,d}) + \hat{B}_m(\hat{t}_{y,d}) \quad (21)$$

où $\hat{V}_{SAM}(\hat{t}_{y,d})$, $\hat{V}_{NR}(\hat{t}_{y,d})$, $\hat{V}_{MIX}(\hat{t}_{y,d})$ et $\hat{B}_m(\hat{t}_{y,d})$ représentent respectivement les estimations de la variance échantillonnale, de la variance due à l'imputation, du terme de covariance et du terme de biais relié au modèle d'imputation. À noter que l'équation (21) représente notre définition de la variance totale si le biais $B_m(\hat{t}_{y,d})$ est négligeable et on aura :

$$\hat{V}_{TOT}(\hat{t}_{y,d}) = \hat{V}_{SAM}(\hat{t}_{y,d}) + \hat{V}_{NR}(\hat{t}_{y,d}) + \hat{V}_{MIX}(\hat{t}_{y,d}) \quad (22)$$

La variance échantillonnale est définie de la manière suivante :

L'estimation de la variance échantillonnale peut être décomposée en 2 ter-

$$\text{mes } \hat{V}_{SAM}(\hat{t}_{y,d}) = \hat{V}_{ORD}(\hat{t}_{y,d}) + \hat{V}_{DIF}(\hat{t}_{y,d})$$

où

$\hat{V}_{ORD}(\hat{t}_{y,d})$ = l'estimateur naïf de la variance en considérant les valeurs imputées comme si elles

avaient été rapportées (pour plus d'information sur le contexte de cet estimateur dans le Pise le lecteur peut se référer à (Turmelle et Beaucage 2013, [15]), et

$$\hat{V}_{DIF}(\hat{t}_{y,d}) = \sum_{k \in S_{my_v}} f_{y_v dk}^2 (1 - \pi_k) w_k^2 \hat{\sigma}_{y_v k}^2$$

est un terme de correction proposé par (Särndal C.-E., 1992, [14]) et simplifié par (Beaumont et Bocci 2009, [6]) pour tenir compte de l'effet de l'imputation sur l'estimation de la variance échantillonnale.

L'estimation de la variance due à l'imputation est donnée par :

$$\hat{V}_{NR}(\hat{t}_{y,d}) = \sum_{k \in s_{ryy}} W_{y,dk}^2 \hat{\sigma}_{y,k}^2 + \sum_{k \in s_{myy}} w_k^2 f_{y,dk}^2 \hat{\sigma}_{y,k}^2 \quad (23)$$

où $W_{y,dk}$ est défini par l'équation (3).

L'estimation du terme de covariance est donnée par :

$$\hat{V}_{MIX}(\hat{t}_{y,d}) = 2 \sum_{k \in s_{ryy}} W_{y,dk} (w_k - 1) f_{y,dk} \hat{\sigma}_{y,k}^2 - 2 \sum_{k \in s_{myy}} w_k (w_k - 1) f_{y,dk}^2 \hat{\sigma}_{y,k}^2 \quad (24)$$

où $W_{y,dk}$ est défini par l'équation (3).

Finalement, le terme de biais relié au modèle est donné par :

$$\hat{B}_m(\hat{t}_{y,d}) = W_{0y,d} + \sum_{k \in s_{ryy}} W_{y,dk} \hat{\mu}_{y,k} - \sum_{k \in s_{myy}} w_k f_{y,dk} \hat{\mu}_{y,k} \quad (25)$$

où $W_{0y,d}$ et $W_{y,dk}$ sont définis par l'équation (3).

Pour plus de détails concernant la variance due à la non-réponse et $\hat{V}_{DIF}(\hat{t}_{y,d})$, le lecteur peut se référer à l'article (Beaumont & Bissonnette 2011, [4]).

3.4.3.2. Impact (MI) sur les différentes composantes de la variance totale

L'impact sur les différentes composantes de la variance pour une unité k sous certaines hypothèses (décrites en annexe) sont:

Impact sur $\hat{V}_{ORD}(\hat{t}_{y,d})$ du traitement T sur l'unité convertie k

$$\delta_k(\hat{V}_{ORD}(\hat{t}_{y,d})) = \hat{V}_{ORD}(\hat{t}_{y,d}) - \hat{V}_{ORD}^{(k)}(\hat{t}_{y,d}) = 0$$

Impact sur $\hat{V}_{DIF}(\hat{t}_{y,d})$ du traitement T sur l'unité convertie k

$$\begin{aligned} \delta_k(\hat{V}_{DIF}(\hat{t}_{y,d})) &= \hat{V}_{DIF}(\hat{t}_{y,d}) - \hat{V}_{DIF}^{(k)}(\hat{t}_{y,d}) \\ &= f_{y,dk}^2 (1 - \pi_k) w_k^2 \hat{\sigma}_{y,k}^2 \end{aligned}$$

Impact sur $\hat{V}_{NR}(\hat{t}_{y,d})$ du traitement T sur l'unité convertie k

$$\begin{aligned} \delta_k(\hat{V}_{NR}(\hat{t}_{y,d})) &= \hat{V}_{NR}(\hat{t}_{y,d}) - \hat{V}_{NR}^{(k)}(\hat{t}_{y,d}) \\ &= \sum_{i \in s_{ryy}} (2W_{y,di} w_k f_{y,dk} \varphi_{y,ik} - w_k^2 f_{y,dk}^2 \varphi_{y,ik}^2) \hat{\sigma}_{y,i}^2 + w_k^2 f_{y,dk}^2 \hat{\sigma}_{y,k}^2 \end{aligned}$$

Impact sur $\hat{V}_{MIX}(\hat{t}_{y,d})$ du traitement T sur l'unité convertie k

$$\begin{aligned} \delta_k(\hat{V}_{MIX}(\hat{t}_{y,d})) &= \hat{V}_{MIX}(\hat{t}_{y,d}) - \hat{V}_{MIX}^{(k)}(\hat{t}_{y,d}) \\ &= 2 \sum_{i \in s_{ryy}} w_k f_{y,dk} \varphi_{y,ik} (w_i - 1) f_{y,di} \hat{\sigma}_{y,i}^2 - 2w_k (w_k - 1) f_{y,dk}^2 \hat{\sigma}_{y,k}^2 \end{aligned}$$

Impact sur $\hat{V}_{TOT}(\hat{t}_{y,d})$ du traitement T sur l'unité convertie k

$$\begin{aligned}\delta_k(\hat{V}_{TOT}(\hat{t}_{y,d})) &= \hat{V}_{TOT}(\hat{t}_{y,d}) - \hat{V}_{TOT}^{(k)}(\hat{t}_{y,d}) \\ &= \delta_k(\hat{V}_{ORD}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{DIF}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{NR}(\hat{t}_{y,d})) + \delta_k(\hat{V}_{MIX}(\hat{t}_{y,d}))\end{aligned}$$

Pour une description plus détaillée des hypothèses et du développement des termes d'impact décrits dans cette section, le lecteur est référé à (Bosa et coll. 2018, [8]).

4. Mise en œuvre de la stratégie dans le Pise

L'IQMI a été progressivement mis en œuvre pour les enquêtes du Pise depuis 2015 :

- IQMI a été utilisé pour la première fois auprès de trois enquêtes en 2015, pour une période de six semaines seulement à la fin de la collecte active, et a été testé après la collecte pour 20 autres enquêtes. Dans tous ces cas, IQMI n'a utilisé seulement que le taux de réponse pondéré de la variable clé. Outre la période où IQMI a été utilisé, toutes les enquêtes ont utilisé pendant la collecte une liste de priorisation statique, créée avant le début de la collecte. Cette liste visait un compromis de qualité pour les variables clés, et a été créée en utilisant des valeurs prédites. Pour ceci, une approche semblable au calcul des MI globales a été utilisée, mais seulement en tenant compte des estimations clés au niveau national.
- En 2016 IQMI a été utilisé pour environ 30 enquêtes annuelles et une enquête mensuelle, en utilisant les taux de réponse pondérés de la variable clé et ERVP (bien que l'IQ utilisé diffère de celui présenté à la section 3.4.2, basé sur un écart au carré plutôt qu'absolu). Pour les enquêtes annuelles, IQMI a débuté environ trois mois après le début de la période de collecte. Avant IQMI, toutes les enquêtes ont utilisé pour les opérations de suivi une liste de priorisation statique selon une stratégie semblable à 2015, mais en considérant cette fois toutes les estimations clés des enquêtes.
- En 2017 IQMI a été utilisé pour environ 45 enquêtes, toujours selon ces deux indicateurs de qualité, en utilisant cette fois l'IQ d'ERVP décrit à la section 3.4.2. Pour les enquêtes annuelles correspondantes à celles utilisant IQMI en 2016, IQMI a été introduit un mois après le début de la collecte (soit essentiellement pour le début des suivis de la collecte) et ces enquêtes ont utilisé avant IQMI une stratégie semblable à celle utilisée en 2016.
- En 2018 il est prévu d'utiliser IQMI pour environ 65 enquêtes, selon ces deux indicateurs de qualité, en débutant IQMI au début des opérations de suivi de collecte. Il est aussi prévu de tester l'utilisation de l'indicateur de qualité selon le CV et selon des taux de réponse pondérés et non pondérés (voir les cas particuliers de la section 3.4.1), en prévision d'une mise en œuvre future pour plusieurs enquêtes.

Le lecteur est invité à consulter (Andrews et coll. 2016, [1]), (Mireuta et coll. 2017, [12]) et (Andrews et coll. 2018, [2]) pour plus d'information sur la mise en œuvre de IQMI pour les enquêtes du Pise.

5. Impact de IQMI sur les enquêtes du Pise

5.1. Conceptuel

L'utilisation d'IQMI pour les enquêtes du Pise implique un changement notable à la façon de mesurer la qualité pendant la période de collecte. La mesure de qualité standard précédemment utilisée par la

majorité des enquêtes entreprises pour décider du succès des opérations de collecte était un indicateur selon un taux de réponse simple pondéré, souvent dérivé selon le revenu opérationnel multiplié par le poids de sondage, et utilisait un nombre restreint de domaines. IQMI, tel qu'expliqué dans les sections précédentes, permet d'évaluer la qualité d'une multitude d'estimateurs selon plusieurs mesures de qualité. La plupart des enquêtes comprennent des estimations clés définies à partir de quatre à six variables importantes et de domaines se basant sur quelques variables de classification (par exemple géographie et industrie). De cette façon un nombre parfois considérable d'estimations clés sont surveillées pendant la période de collecte des enquêtes, et IQMI priorise les suivis de collecte pour l'ensemble de ses estimations. Une fois que la cible de qualité d'une estimation est atteinte les efforts de collecte sont redirigés vers les unités qui vont le plus améliorer la qualité des autres domaines n'ayant pas atteint leur cible. Ainsi, l'emphase n'est plus sur une seule variable, par exemple le revenu, et les domaines d'intérêt correspondent mieux aux objectifs de l'enquête. Dans ce contexte, il est normal de s'attendre à une amélioration générale de la qualité des données pour les enquêtes, ce qui améliore leur pertinence pour les usagers.

L'information produite par le processus d'IQMI est aussi progressivement utilisée par les agents spécialisés pour l'analyse de données. Elle leur permet d'identifier efficacement les unités qui ont le plus d'influence sur la qualité des estimations, et de débiter tôt les vérifications des cas les plus problématiques qui ne sont plus éligibles aux suivis de collecte. Ils peuvent également identifier les estimations clés pour lesquelles les cibles de qualité ont été atteintes, et ainsi débiter l'analyse de données plus rapidement.

De plus, le processus automatisé d'IQMI rend la gestion active de la collecte et de l'analyse beaucoup plus efficace pour détecter et s'ajuster aux problèmes de qualité pour les enquêtes que les processus utilisés précédemment, qu'il soit utilisé en remplacement ou en complément de ceux-ci.

5.2. Taux de réponse

Avec l'utilisation d'IQMI, une inquiétude initiale des enquêtes était que la priorisation ajustée pour augmenter l'emphase mise sur des domaines plus petits pourrait avoir un impact négatif sur les taux de réponse simples pondérés (selon le revenu) et nuire à la collecte d'unités importantes pour les domaines nationaux. Pour la très grande majorité des enquêtes évaluées ayant utilisé IQMI, une amélioration notable a été remarquée non seulement pour la qualité de l'ensemble des estimations clés, mais aussi pour les taux de réponse simples utilisés précédemment pour évaluer le succès de la collecte. Pour les deux enquêtes évaluées où les résultats avec IQMI n'étaient pas meilleurs en 2017, les taux de réponse simples étaient très proches. Une analyse de l'impact sur les enquêtes ayant utilisé IQMI a été complétée pour trois années de collecte, de 2015 à 2017. Le contexte de l'utilisation d'IQMI pour ces enquêtes annuelles a été présenté à la section 4. Quelques faits saillants de cette étude sont présentés dans les figures 1 à 4. Pour les figures 1 à 3, associées à une enquête représentative des enquêtes étudiées, le point de départ des courbes correspond au début des estimations en continu pour l'année correspondante, et les courbes sont en pointillées pour les périodes où IQMI n'a pas été utilisé.

La figure 1 compare les taux de réponse simples pondérés (selon le revenu) de 2015 à 2017, selon les données extraites de l'opération de collecte, pour l'enquête annuelle sur les industries de service : design spécialisé. Le graphique démontre clairement que les taux de réponses simples pondérés sont significativement meilleurs pour les années où IQMI fut utilisé. En 2017, la période de collecte a été interrompue plus tôt pour cette enquête et le taux de réponse est demeuré plus élevé que noté pour

2016. Des résultats similaires ont été obtenus pour le reste des enquêtes étudiées, avec quelques variations en ce qui concerne les légères différences entre 2016 et 2017. Il faut noter que cette enquête fait partie d'un groupe d'enquêtes qui partagent les mêmes ressources de collecte, et il y a ainsi des semaines où certaines de ses enquêtes ont reçu plus ou moins d'efforts de collecte dépendamment de leur statut de collecte.

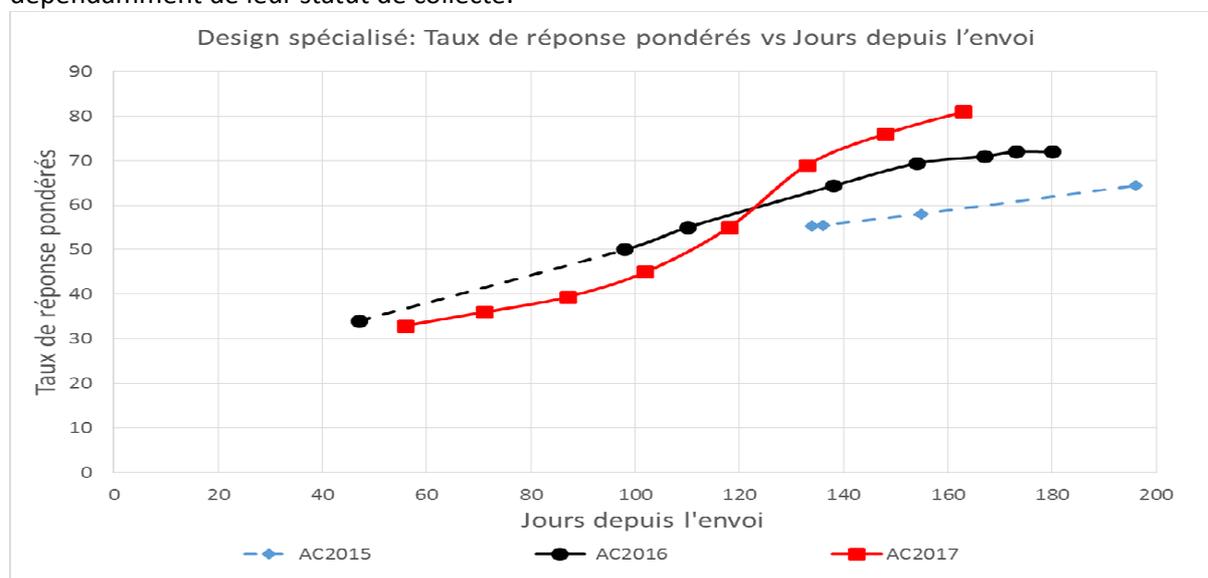


Figure 1. Comparaison des taux nationaux de réponse simples pondérés (selon le revenu) pour les années de collecte (AC) 2015 à 2017

La figure 2 compare le pourcentage de cibles de qualité pour le taux de réponse pondéré selon la variable clé qui ont été atteintes pour les trois mêmes années de collecte, selon les données produites par les estimations en continu. Lors de l'année de collecte 2015, IQMI n'a pas été utilisé et moins de 60% des cibles (celles utilisées en 2016) ont été atteintes. En 2016 l'IQMI n'a été utilisé par la collecte que trois mois après le début des opérations, et malgré ceci le pourcentage est significativement plus élevé qu'obtenu en 2015. En 2017, les cibles de qualité ont été augmentées pour cette enquête, donc bien que le pourcentage obtenu ait été un peu moins élevé à la fin de la collecte active, la qualité obtenue a été légèrement supérieure et a été atteinte plus rapidement.

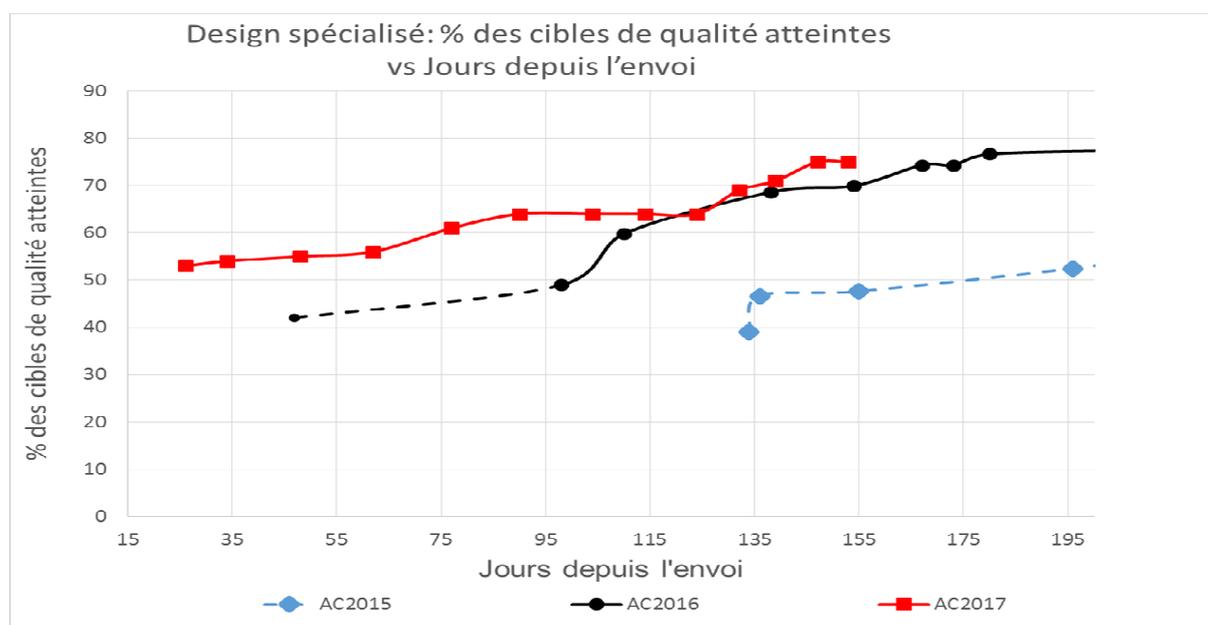


Figure 2. Comparaison du pourcentage de cibles de qualité atteintes pour les années de collecte 2015 à 2017

Pour la figure 3, on peut voir qu'en termes d'effort cumulé des intervieweurs pour les suivis (en minutes) il y a eu une amélioration marquée de la qualité des estimations pour les années où IQMI a été utilisé. On rappelle que la période de collecte active de cette enquête s'est terminée plus tôt en 2017 et que les cibles de qualité ont été majorées comparativement à l'année précédente. Donc, les améliorations en qualité pour les enquêtes ne sont pas associées à un surplus d'effort et le graphique démontre que le niveau de qualité atteint en 2015 aurait pu être obtenu pour les années subséquentes en investissant moins d'effort pour la majorité des enquêtes.

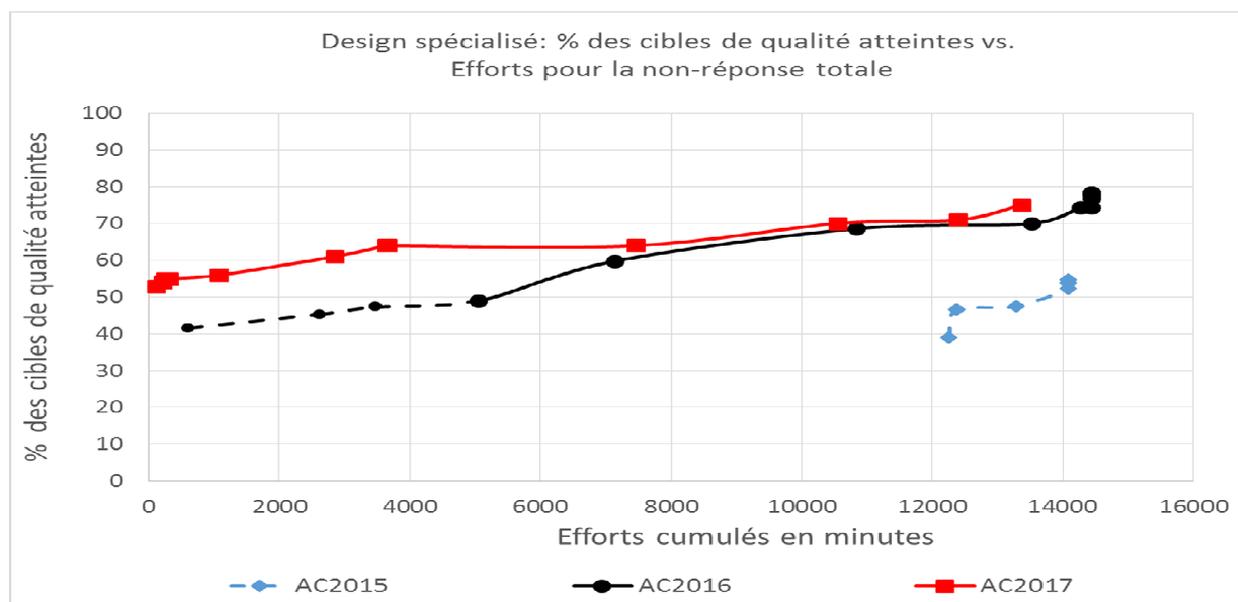


Figure 3. Comparaison des cibles de qualité atteintes versus les efforts cumulés de suivis pour la non-réponse déployés pour les années de collecte 2015 et 2017.

Finalement, la figure 4 compare la moyenne des DQ locales pour 25 enquêtes où des données détaillées étaient disponibles pour une date proche de la fin de la collecte active. La DQ locale a été calculée pour la mesure de qualité du taux de réponse pondéré par la variable clé. Pour 2015, les cibles de qualité et estimations clés de 2016 ont été utilisées pour dériver les DQ locales. Pour toutes sauf trois des enquêtes cette valeur moyenne a été inférieure en 2017 qu'en 2015, même si pour plusieurs enquêtes les cibles de qualité étaient supérieures en 2017 que celles utilisées pour les calculs de 2015 et que la période de collecte s'est terminée plus tôt.

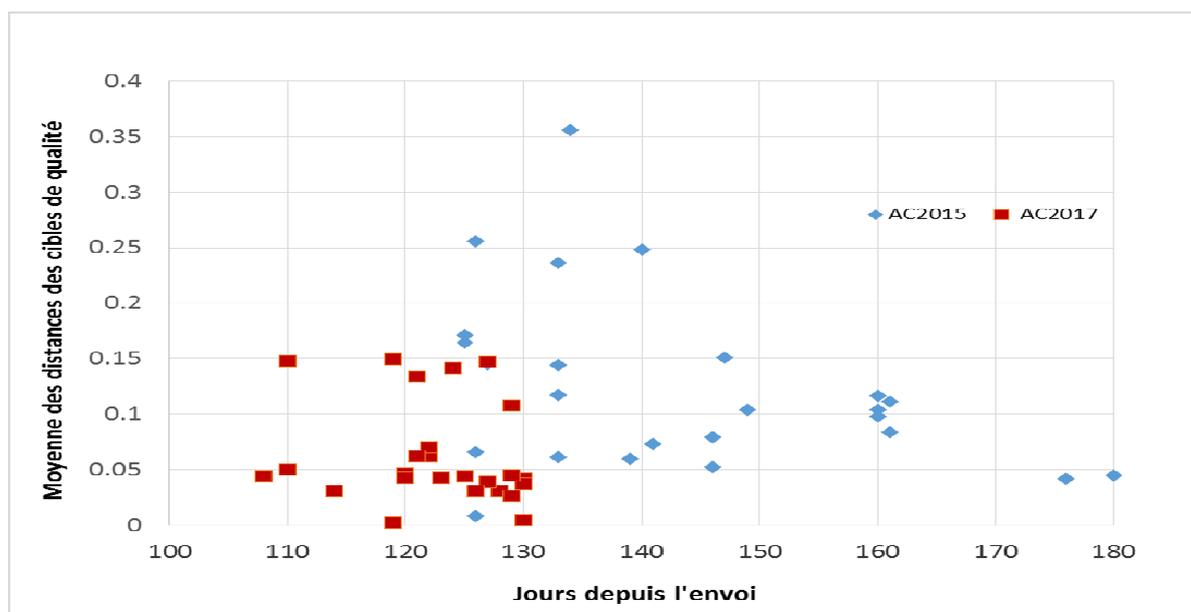


Figure 4. Comparaison de la moyenne des DQ locales pour le taux de réponse pondéré par la variable clé pour 25 enquêtes et les années de collecte 2015 et 2017.

On devrait cependant noter qu'il est possible que les améliorations notées dans cette section soient reliées en partie à la plus grande familiarité des répondants avec la collecte électronique, ou à d'autres facteurs reliés à la collecte, et pas seulement aux améliorations de la priorisation selon IQMI.

5.3. Actualité

Une autre préoccupation pressante en ce qui concerne la qualité à Statistique Canada est l'actualité des données produites. Pour le groupe de 36 enquêtes considérées pour cette étude, toutes sauf une ont publié plus de douze mois après la fin de l'année de référence pour les données recueillies en 2015. Pour ces enquêtes, il est souhaitable que les données soient publiées moins d'un an après la fin de l'année de référence et une étude a été entreprise visant à réduire la période de collecte en 2016, et encore plus significativement en 2017, pour atteindre cet objectif. Pour les données recueillies en 2017, un tiers de ces enquêtes ont été en mesure d'atteindre cet objectif. L'amélioration de la priorisation des unités pour les suivis reliés à IQMI et une réallocation des efforts de collecte ont mené à une réduction importante de la période de collecte active pour ces enquêtes. Tel qu'indiqué à la section précédente, la qualité atteinte pour les enquêtes n'a pas souffert de cette réduction de la période de collecte, et plusieurs enquêtes auraient même pu terminer la collecte active encore plus tôt et atteindre le niveau de qualité de l'année précédente. Par conséquent, l'objectif pour les années futures est de poursuivre la réduction des périodes de collecte lorsqu'un pourcentage suffisant des cibles de qualité aura été atteint ou encore lorsque les distances de qualité seront suffisamment petites. Cette réduction de la période de collecte, et possiblement des efforts de collecte, améliorera non seulement l'actualité des données publiées mais diminuera également les coûts.

6. Plans futurs

L'intégration d'IQMI aux enquêtes du Pise se poursuivra au cours des prochaines années. Les améliorations méthodologiques et de performance apportées au système, l'amélioration des connaissances générales sur la méthodologie de l'IQMI et de son impact sur les enquêtes, et de façon générale l'expérience acquise par tous les intervenants au cours des dernières années, accélèrent grandement la courbe d'apprentissage et la mise en œuvre de cette approche pour les enquêtes qui s'intègrent au Pise.

Au cours des prochaines années l'IQMI débutera l'utilisation de quelques indicateurs de qualité additionnels. Des taux de réponses simples seront utilisés pour quelques enquêtes qui utilisent leurs données pour la création de bases de sondage et pour la modélisation, et les indicateurs selon le CV basé sur la variance totale seront progressivement utilisés par la majorité des enquêtes du Pise. L'ajout de ces indicateurs de qualité pourrait modifier les priorités de collecte pour les enquêtes qui utilisent déjà IQMI. Par exemple, avec l'ajout de l'indicateur selon le CV, des unités plus petites mais pour lesquelles l'imputation est plus variable pourraient recevoir des scores de mesure d'impact final plus élevés puisque ceci pourrait mener à une amélioration des CV des estimations de l'enquête. En contrepartie des unités qui avaient un impact important sur les taux de réponse par exemple, mais pour lesquels des valeurs imputées sont considérées précises pourraient voir une diminution de leur priorité.

Finalement, des travaux se poursuivront pour la modélisation à la fois de la propension de réponse sous un traitement T ou sous l'absence d'un traitement T, ainsi qu'au développement à la méthodologie actuelle pour tenir compte de ceci. En particulier, ces travaux reposent sur l'utilisation de données massives qui proviennent de parodonnées de collecte, selon les différents contacts avec les répondants et selon les logs de questionnaires électroniques, et sur d'autres données auxiliaires connexes. Ce projet cadre dans l'initiative de modernisation en cours présentement à Statistique Canada, et implique des connaissances du domaine des sciences de données (données volumineuses, apprentissage machine). Cette recherche vise à augmenter ou diminuer les priorités (des suivis de collecte principalement) selon la situation contextuelle du répondant. Un répondant qui a reçu plus de contacts déjà sans succès ou encore pour lequel on détecte une grande activité pour le questionnaire électronique pourrait avoir une légère réduction de sa priorité, alors qu'un répondant qui semble stagner dans son cheminement pour la collecte électronique pourrait voir sa priorité augmenter.

7. Conclusion

Cet article a décrit l'utilisation pour le Pise d'une méthodologie de gestion active de la collecte et de l'analyse en se fondant sur des indicateurs de qualité et le calcul de scores selon une mesure d'impact visant à améliorer la qualité. Il a également été démontré que, pour les premières années de l'utilisation, cette approche a eu un impact positif sur les enquêtes du Pise. Elle a contribué à améliorer le taux de réponse simple pondéré selon le revenu au niveau national ainsi que la qualité de l'ensemble des estimations clés des enquêtes, et a permis de publier plus rapidement les données pertinentes aux usagers. Le développement d'IQMI se poursuivra au cours des prochaines années, ce qui devrait offrir une meilleure mesure de la qualité lors des activités de priorisation, et mener à des gains supplémentaires sur la qualité des données ainsi qu'à une réduction des coûts des enquêtes.

8. Reconnaissances

Les auteurs aimeraient remercier Serge Godbout et Jean-François Naud pour la revue de cet article et leurs nombreux commentaires utiles.

Bibliographie

- [1] Andrews, J., Daoust, P. et Mireuta M. (2016). « Small Domains and Sparse Variables: Challenges with creating an Active Collection Strategy ». Recueil de la section des méthodes d'enquêtes, Rencontre annuelle de la SSC, juin 2016.
- [2] Andrews, J., Mireuta, M., Muhirwa, L., et Daoust, P. (2018). « Increasing Data Quality for Business Surveys without Impacting Collection Effort ». À paraître dans Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference, mai 2018.

- [3] Bankier, M. (1988). « Power allocations: Determining sample Sizes for subnational areas ». *The American Statistician*, Vol. 42, 174-177.
- [4] Beaumont, J-F. et Bissonnette, J. (2011). « Estimation de la variance sous imputation composite : méthodologie programmée dans le Sevani ». *Technique d'enquête*, Vol. 37, 183-192.
- [5] Beaumont, J-F., Bissonnette, J. et Bocci, C.(2010). « Sevani, version 2.3, Methodology Guide ». Internal report, Methodology Branch, Statistics Canada.
- [6] Beaumont, J.-F., et Bocci, C. (2009). « Variance Estimation when Donor Imputation is Used to Fill in Missing Values ». *Canadian Journal of Statistics*, 37, 400-416.
- [7] Bosa, K. et Godbout, S. (2014), « Mesures de qualité du PISE – Guide méthodologique – », Division des méthodes d'enquêtes entreprises, document interne de Statistique Canada, révisé novembre 2014.
- [8] Bosa, K., Godbout, S., Mills, F., et Picard F. (2018). «How to Decompose the Nonresponse Variance: A Total Survey Error Approach ». À paraître dans *Technique d'enquête*, 2018.
- [9] Godbout, S. (2011). « Normalisation du traitement des données après la collecte dans les enquêtes-entreprises à Statistique Canada ». *Recueil du Symposium international de 2011 sur la méthodologie*.
- [10] Godbout, S., Beaucage, Y. et Turmelle, C. (2011). « Quality and Efficiency using a Top-Down Approach in the Canadian's Integrated Business Statistics Program ». *Conference of European Statisticians, Work Session on Statistical Data Editing. Ljubljana, Slovénie (9-11 mai 2011)*.
- [11] Hedlin, D. (2008). « Local and Global Score Functions in Selective Editing ». *United Nations Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, Work Session on Statistical Data Editing, Vienne, Autriche 2008*.
- [12] Mireuta, M., Andrews, J., et Daoust, P. (2017). « The Relative Deviation from Predicted Values as a Tool to Prioritize Units for Failed Edit Follow-Up in IBSP ». *Recueil de la section des méthodes d'enquêtes, Rencontre annuelle de la SSC, juin 2017*.
- [13] Mills, F., Godbout, S., Bosa, K. and Turmelle, C. (2013), « Multivariate Selective Editing in the Integrated Business Statistics Program », *Proceedings of the Joint Statistical Meeting 2013, American Statistical Association*.
- [14] Särndal, C.E., Swensson, B. and Wretman, J. (1992). « *Model Assisted Survey Sampling* ». New York, Springer-Verlag, 694 p.
- [15] Turmelle, C. and Beaucage, Y. (2013), « Le Programme intégré de la statistique des entreprises : Utilisation d'un plan de sondage à deux phases pour produire des estimations fiables », *Recueil du Symposium international de 2013 sur la méthodologie*.
- [16] Turmelle, C., Godbout, S. et Bosa, K. (2012). « Methodological Challenges in the Development of Statistics Canada's New Integrated Business Statistics Program ». *Proceedings of the ICES IV Conference*.
- [17] Turmelle, C., Godbout, S., Bosa, K. et Mills, F. (2014) « A Quality Driven Approach to Managing Collection and Analysis », *Proceedings of the European Conference on Quality in official statistics. Vienne*.
- [18] Statistique Canada (2015), « Aperçu du Programme intégré de la statistique des entreprises », *No 68-515-X au catalogue*, diffusé le 17 juin 2015.

Annexe: Hypothèses utilisées pour dériver les MIs reliées à la variance totale

Certaines hypothèses doivent être faites pour déterminer l'impact (MI) qu'a une unité sur les différentes composantes de la variance totale, présentées à la 3.4.3.2. Dans un premier temps, les unités répondantes $k \in s_{ry_v}$ auront une MI=0 associée au CV, ces unités sont répondantes et ne contribuent pas à la variance due à la non-réponse. Les différentes hypothèses qui sont utilisées pour dériver les mesures d'impact de la section 3.4.3.2 sont décrites ci-dessous.

Hypothèses :

1. On suppose qu'on a une unité $k \in s_{my_v}$ qui contribue à un domaine d et que cette dernière sera convertie dans s_{ry_v} suite au suivi de collecte; on utilisera l'exposant (k) pour exprimer que l'unité k est maintenant considérée dans l'ensemble s_{ry_v} . Cette notation (exposant (k)) a été introduite à la section 3.3.1.
2. On suppose que $y_{vk}^{(I)} = y_{vk}^{(R)}$ ou que ces deux valeurs sont très proches
3. On suppose que l'estimation des paramètres dans le modèle d'imputation demeure les mêmes une fois l'unité k convertie en répondante, c.-à-d.

- $\hat{\mu}_{y,k}$ demeure inchangé
- $\hat{\sigma}_{y,k}^2$ demeure inchangé

4. On réduit le poids de compensation $W_{y,dk}$ que porte l'unité répondante k' par la contribution de l'unité convertie k . On aura donc que
$$W_{y,dk}^{(k)} = \sum_{i \in s_{my_v}} w_i f_{y,di} \phi_{y,k'i} - w_k f_{y,dk} \phi_{y,k'k} = W_{y,dk} - w_k f_{y,dk} \phi_{y,k'k} \quad \forall k' \in s_{ry_v}$$
 sous l'hypothèse 1.

À noter que sous l'hypothèse 1, k est maintenant une unité répondante et sa nouvelle valeur $\phi_{y,k'k}^{(k)}$ devrait être 0 si on recalculait les nouveaux termes $W_{y,dk}$.

5. On suppose que l'unité k convertie n'est pas utilisée pour imputer les unités non répondantes, ce qui signifie que $\phi_{y,ki}^{(k)} = 0 \quad \forall i \in s_{my_v}$ et donc que
$$W_{y,dk}^{(k)} = \sum_{i \neq k \in s_{my_v}} w_i f_{y,di} \phi_{y,ki}^{(k)} = 0.$$