N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2004 : Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

2004

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2004 de Statistique Canada
Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre

LES DÉFIS RELIÉS À L'ENQUÊTE SUR LE COMMERCE ÉLECTRONIQUE ET LA TECHNOLOGIE

Marie-Claude Duval et Edith Latendresse[1]

RÉSUMÉ

L'Enquête sur le commerce électronique et la technologie (ECET) permet d'obtenir de l'information sur les technologies de l'information et des communications telles l'utilisation d'Internet, de courriels, d'Intranet, d'un site Web, ainsi que l'utilisation ou non d'Internet à des fins de commerce électronique. Il s'agit d'une enquête annuelle réalisée auprès d'un échantillon de 20 000 entreprises canadiennes et couvrant tous les secteurs industriels de l'économie. Depuis sa création en 1999, les besoins en données se sont accrus et l'expertise reliée au domaine de la technologie informatique et du commerce électronique s'est développée. Ceci a entraîné un questionnement sur la méthodologie de l'enquête utilisée, construite sur un modèle d'enquête économique standard. Jusqu'à maintenant, de nombreuses études portant sur différents aspects ont été réalisées dont la couverture de la population observée, le plan d'échantillonnage, la collecte ainsi que la vérification et l'imputation. D'autres études devront aussi être faites concernant l'estimation d'un événement rare. Le présent article présentera l'enquête, ses principaux enjeux et défis, des résultats d'études ainsi que les mesures prises jusqu'à présent et celles à venir.

MOTS CLÉS: Enquête économique, événement rare, plan d'échantillonnage, variables dichotomiques.

1. INTRODUCTION

L'Enquête sur le commerce électronique et la technologie est une enquête annuelle qui recueille de l'information sur l'utilisation de la technologie et le commerce électronique auprès des entreprises canadiennes de tous les secteurs industriels. Les principales variables d'intérêt sont l'utilisation d'Internet, la possession d'un site Web ainsi que l'achat et /ou la vente de biens et services via Internet. On demande également les contraintes du commerce électronique pour les entreprises ne faisant pas d'achat ou de vente par Internet ainsi que les améliorations technologiques des entreprises au fil des années. Cette enquête existe depuis 1999.

Contrairement à une enquête économique standard où on mesure généralement des variables économiques numériques telles les composantes de dépenses et de revenus, l'enquête sur le commerce électronique mesure plutôt des variables qualitatives, plus spécifiquement des variables dichotomiques. On s'intéresse à l'utilisation ou non d'une caractéristique. Par exemple, utilisez-vous Internet (oui/non)? Nous avons constaté au fil des années certaines lacunes du plan d'échantillonnage actuel étant donné la nature de l'enquête, le type de variables mesurées et les besoins d'analyse.

On mesure également une seule variable numérique mais très importante à l'enquête, soient les ventes totales réalisées via le commerce électronique. Depuis le début de l'enquête en 1999, on a plutôt opté pour un plan d'échantillonnage qui permettait de mesurer le mieux possible les ventes totales par Internet selon la base de sondage qui nous était disponible. Or, nous avons réalisé l'instabilité de cette variable dans le temps malgré le plan d'échantillonnage utilisé. Nous sommes, d'une part, en présence d'un évènement rare. Peu d'entreprises font du commerce électronique et certains secteurs industriels n'ont pas un marché destiné aux ventes par Internet. D'autre part, nous avons une variable dont les concepts ne semblent pas toujours bien compris par les répondants ou encore dont l'information n'est pas disponible. Par exemple, nous avons noté que certaines entreprises faisant des ventes par Internet connaissent leurs ventes totales, mais ne savent pas spécifiquement celles réalisées via Internet.

[1]Marie-Claude Duval, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11e étage, immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6 (marie-claude.duval@statcan.ca), Edith Latendresse, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11e étage, immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6 (edith.latendresse@statcan.ca)

Le présent article décrit la méthodologie utilisée pour l'enquête basée sur une enquête économique standard, la réévaluation des besoins au fil des années, les lacunes rencontrées, la méthodologie proposée afin de mieux répondre aux besoins et finalement les défis à venir.

2. MÉTHODOLOGIE DE L'ENQUÊTE

L'enquête actuelle est basée sur un plan d'échantillonnage optimum pour mesurer des variables reliées au revenu de l'entreprise, plan couramment utilisé dans les enquêtes économiques à Statistique Canada. Il s'agit essentiellement de minimiser la taille de l'échantillon pour un niveau de précision désiré dans les estimations. La section qui suit présentera tout d'abord la population couverte par l'enquête, ensuite la stratification et la répartition de l'échantillon, la sélection de l'échantillon, la vérification et l'imputation et, finalement, l'estimation.

La population d'entreprises est tirée du registre des entreprises (RE) de Statistique Canada (Cuthill, 1997) qui compte près de 2 millions d'entreprises. La première étape consiste à exclure de la population les plus petites unités. On considère ces unités comme ayant peu ou pas d'impact sur l'activité économique. On définit les bornes d'exclusion par domaine d'intérêt de sorte à exclure au maximum 5% du revenu total par domaine d'intérêt. Le revenu utilisé est le revenu brut de l'entreprise disponible pour toutes les unités du RE. On obtient alors une population observée d'environ 650 000 unités qui est également notre population cible. Les domaines d'intérêt sont définis selon la classification industrielle en Amérique du Nord (SCIAN) à différents niveaux de détail selon le secteur industriel.

La population observée est ensuite stratifiée par taille à l'intérieur des domaines SCIAN en utilisant l'algorithme de Lavallée-Hidiroglou (1988). L'algorithme permet de grouper les unités en strates à tirage partiel et strate à tirage complet et de répartir la taille de l'échantillon par strate de sorte à minimiser la taille d'échantillon pour un coefficient de variation (CV) donné. Nous spécifions 2 strates à tirage partiel et 1 strate à tirage complet par domaine d'intérêt. Pour calculer le CV, nous utilisons comme variable auxiliaire le revenu brut de l'entreprise disponible pour toutes les unités sur le RE. Dans le cadre de l'enquête, le budget disponible d'environ 20 000 unités est connu. C'est pourquoi notre approche est plutôt de tester différents CV et de choisir le CV qui respecte la taille d'échantillon permise. Nous voulons également un plus grand échantillon dans les secteurs plus propices au commerce électronique afin d'améliorer la qualité des estimations des ventes totales par Internet. Pour ce faire, en 2003, nous avons utilisé un CV cible de seulement 2% dans les secteurs de la fabrication, du commerce de gros, du commerce de détail, du transport et de l'entreposage, plus susceptibles de faire du commerce électronique et un CV cible de 5% dans tous les autres secteurs industriels. Finalement, nous augmentons la taille de l'échantillon dans certaines strates, si nécessaire, de sorte à obtenir une fraction d'échantillonnage minimale de 1% et un minimum de cinq unités échantillonnées par strate.

Toutes les unités dans les strates à tirage complet sont échantillonnées avec certitude. De plus, quelques unités pour lesquelles on s'attend à de très grandes ventes par Internet sont identifiées et forcées dans l'échantillon avec un poids de sondage de 1. Un échantillon aléatoire est tiré dans les strates à tirage partiel sous la contrainte de maximiser le chevauchement avec l'échantillon de l'année précédente. La méthode de Kish et Scott (1971) est utilisée et en 2003, un chevauchement global de 65% a été obtenu avec l'échantillon précédent. Chaque année, une nouvelle population est créée et un nouvel échantillon est tiré.

Une fois l'échantillon tiré, un questionnaire est envoyé par la poste aux entreprises. À la saisie des données, des règles de vérification sont appliquées à chaque questionnaire, telles des règles de cohérence et de vérification historique. Les unités n'ayant pas répondu font l'objet de suivis postaux et par télécopieur afin d'obtenir leurs réponses. Certains suivis par téléphone sont également faits afin d'améliorer le taux de réponse, d'obtenir des données non rapportées ou encore de valider/corriger des données extrêmes ou incohérentes fournis par le répondant. Enfin, nous effectuons les suivis en donnant priorité à certaines unités selon le taux de réponse par strate, le secteur industriel, la taille de l'entreprise, l'importance des variables manquantes et le type d'incohérences sur le questionnaire. Le taux de réponse est d'environ 70% à chaque année.

Une fois la collecte terminée, les enregistrements toujours incomplets et/ou incohérents sont imputés. En moyenne, seulement 6% des champs sont imputés parce que manquants et moins de 0.5% parce qu'incohérents. Différentes méthodes d'imputation sont utilisées telles l'imputation déterministe, l'imputation historique et l'imputation par donneur. Seuls les questionnaires partiels sont imputés. Dans le cas d'une non-réponse totale, aucune imputation n'est faite. On repondère plutôt les unités répondantes. De plus, on fait la détection des données aberrantes en utilisant une variante de la règle de l'écart-sigma (Nobrega, 1998) pour les ventes totales par Internet.

Le système généralisé d'estimation (SGE) de Statistique Canada est utilisé pour faire l'estimation de toutes les variables. Pour de l'information sur le SGE, se référer à Estevao (2005). Deux principaux types d'estimations sont produits :

1) Dans le cas des **variables dichotomiques (*C*)**, un quotient est utilisé.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \quad \text{où } c_i(d) = \begin{cases} 1 \; si \; i \varepsilon \; d \; et \; i \; a \; la \; caractéristique \\ 0 \; sinon \end{cases}$$

2) Dans le cas de la **variable numérique (*Y*)**, un estimateur de total est utilisé.

$$\hat{Y}_d = \sum_s w_i y_i(d) \quad \text{où } y_i(d) = \begin{cases} y_i \; si \; i \varepsilon \; d \\ 0 \; sinon \end{cases}$$

La variable w_i représente le poids de sondage ajusté de l'unité *i* après repondération pour tenir compte de la non-réponse. La variable z_i est une variable auxiliaire qui peut être le revenu, le nombre d'employés ou autre selon la variable estimée. Des estimations sont produites avec et sans cette variable auxiliaire. Cette variable permet de dériver des estimations qu'on appelle économiquement pondérées en donnant plus de poids aux unités de grandes tailles. Par exemple, en 2003, on a estimé à 34% le pourcentage d'entreprises ayant un site Web et 85% lorsque économiquement pondéré selon le revenu. Cela signifie que les 34% d'entreprises ayant un site Web représentent 85% du revenu total au Canada.

Finalement, le SGE calcule une variance et un CV pour toutes les variables estimées.

3. RÉÉVALUATION DES BESOINS

Au fil des années, les besoins en données se sont accrus et une expertise en la matière s'est développée.

Premièrement, la comparaison des estimations dans le temps est très importante pour les utilisateurs. On veut mesurer les changements et l'évolution de l'utilisation des technologies, mais aussi le comportement des entreprises sur l'utilisation des technologies dans le temps. Par exemples, quelle est l'augmentation du pourcentage d'entreprises ayant un site Web de 2001 à 2003? Comment se comportent en 2003 les entreprises qui utilisaient peu la technologie en 2000?

Deuxièmement, depuis les deux dernières années, on s'intéresse à l'utilisation de la technologie par taille d'entreprises (petites, moyennes et grandes entreprises) définie selon le nombre d'employés. On utilise présentement le nombre d'employés demandé sur le questionnaire et on procède à l'estimation par domaine.

La taille d'une entreprise est définie comme suit (à l'exception du secteur manufacturier) :

Petite entreprise:	entreprise avec moins de 20 employés
Moyenne entreprise :	entreprise de 20 à 99 employés
Grande entreprise :	entreprise de 100 employés et plus

Pour le secteur manufacturier, la définition suivante est utilisée:

Petite entreprise:	entreprise avec moins de 20 employés
Moyenne entreprise :	entreprise de 20 à 499 employés
Grande entreprises:	entreprise de 500 employés et plus

Les analystes s'intéressent à l'utilisation et à l'évolution des technologies dans le temps pour les petites, moyennes et grandes entreprises selon ces définitions.

Troisièmement, une des variables d'intérêt les plus importantes à l'enquête est les ventes totales par Internet. Par contre, il s'agit d'un événement rare et nous avons peu d'information auxiliaire disponible afin de cibler les entreprises faisant des ventes par Internet. En 2003, on estimait à 7% seulement les entreprises du secteur privé faisant des ventes par Internet. Les estimations produites jusqu'à maintenant sont souvent de moindre qualité et nous aimerions trouver des moyens pour améliorer la qualité de cette variable.

4. RÉÉVALUATION DES BESOINS VS LA MÉTHODOLOGIE UTILISÉE

Nous avons mentionné à la section 2 qu'une nouvelle population était créée et un nouvel échantillon tiré à chaque année, en excluant de la population les plus petites unités représentant moins de 5% du revenu total par domaine d'intérêt. Nous avons également mentionné que la stratification était basée sur un algorithme permettant de déterminer les bornes de strates par domaine d'intérêt et de déterminer la répartition de l'échantillon par strate de sorte à minimiser la taille d'échantillon requise pour un CV donné. Cette méthode est une méthode optimum lorsqu'on mesure des totaux corrélés à la variable auxiliaire. Par contre, lorsqu'on mesure des variables dichotomiques et que la probabilité d'avoir ou non une caractéristique varie avec la taille de l'entreprise, cette méthode peut parfois donner quelques résultats non concluants.

D'une part, la population cible peut varier considérablement dans le temps dépendamment des changements de revenu des unités d'une année à l'autre. Par exemple, pour les années 2001 et 2002, la population cible globale a augmenté de 2%, ce qui est plausible. Par contre, pour une industrie en particulier, la population cible a augmenté de 41%. Cette augmentation s'explique par la correction à la baisse du revenu d'une seule unité de 2001 à 2002. Ainsi, même si la population totale était similaire d'une année à l'autre, la population cible a changé drastiquement, car, pour être en mesure de couvrir 95% de la population en terme de revenu, on devait ajouter beaucoup d'entreprises. En terme de revenu, nous couvrons la même population, mais en terme de population, nous couvrons beaucoup plus d'entreprises. Un des problèmes est que l'ajout d'entreprises pour couvrir 95% se fait en ordre décroissant et chaque unité ajoutée est plus petite que la précédente. Comme nous savons que les plus petites entreprises utilisent moins la technologie, on craint que l'augmentation du nombre d'entreprises dans la population cible diminue l'estimation par rapport à l'année précédente. La baisse serait donc expliquée par le changement de la population cible et non par l'évolution réelle de la technologie pour ce secteur, conclusion qui n'est pas souhaitable.

D'autre part, le plan d'échantillonnage actuel ne tient pas compte de la représentativité de l'échantillon par taille selon le nombre d'employés. L'algorithme de Lavallée-Hidiroglou (1988) détermine des bornes de strates arbitraires par domaine d'intérêt et répartit la taille d'échantillon par strate. Sa seule contrainte est de minimiser la taille d'échantillon pour un CV ciblé selon une variable auxiliaire, dans notre cas le revenu brut de l'entreprise. Les bornes de strates et la répartition de l'échantillon vont donc varier dépendamment du domaine, du revenu et également de la population cible. Rien ne garantit la représentativité par taille selon le nombre d'employés. Un problème de représentativité a été détecté dans une industrie en comparant les estimations de 2002 et 2003. Le pourcentage estimé d'entreprises ayant un site Web est passé de 68% à 61%. Dans le domaine de la technologie, les estimations à la baisse nous surprennent un peu puisqu'on s'attend au fil des années à une augmentation de

l'utilisation des technologies dans le temps et non à une diminution. Nous avons donc regardé des résultats par taille d'entreprise (Tableau 1).

Tableau 1 : Estimation du pourcentage d'entreprises ayant un site Web et distribution de la population estimée par taille d'entreprises pour une industrie donnée.

Taille	% d'entreprises ayant un site Web		% de la population estimée	
	2002	2003	2002	2003
Petites	60%	54%	72%	83%
Moyennes	87%	97%	23%	13%
Grandes	84%	98%	4%	4%

On peut remarquer le changement de distribution de la population estimée par taille d'entreprise entre les 2 années. On estime à 83% le pourcentage de petites entreprises en 2003 comparativement à 72% en 2002. On constate également que le pourcentage estimé des petites entreprises ayant un site Web est autour de 54% à 60% chez les petites entreprises comparativement à plus de 85% pour les autres entreprises. Il n'est donc pas surprenant que l'estimation de 2003 soit inférieure à celle de 2002 avec autant de petites entreprises dans la population. Est-ce dû à un échantillon non représentatif pour une des deux années? Est-ce dû à un changement de population suite à des changements de revenus de certaines unités? On se rend compte dans ce cas-ci que notre plan d'échantillonnage n'est peut-être pas très robuste pour ce qu'on désire estimer. Heureusement, après analyse, on a trouvé que de tels cas ne sont pas si fréquents, le nombre d'employés ayant une certaine corrélation avec le revenu.

Bien que le plan actuel comporte des lacunes pour certains types d'analyses, il s'agit tout de même d'un plan optimum lorsqu'on calcule des estimations économiquement pondérées. En effet, l'estimateur économiquement pondéré utilise le revenu pour pondérer la valeur de la caractéristique (0 ou 1). Comme la population cible et le plan d'échantillonnage sont construits selon le revenu et couvrent la même population en terme de revenu (95%) d'une année à l'autre, les estimateurs sont donc cohérents.

Finalement, mentionnons que certaines mesures ont été prises afin d'atténuer les lacunes du plan actuel. Tout d'abord une détection des données aberrantes (Nobrega, 1998) est faite sur le revenu de la population dans le temps. Ainsi, les grands changements de revenu sont vérifiés et corrigés au besoin. Ceci permet de stabiliser la population cible dans le temps. En ce qui concerne la distribution par taille, on s'assure d'avoir un taux de réponse homogène par strate. Contrairement à une enquête économique où l'on recherche plutôt une bonne couverture des unités répondantes en terme de revenu ou autre variable auxiliaire, on recherche plutôt dans cette enquête-ci un bon taux de réponse autant chez les entreprises à petits revenus que chez les entreprises à grands revenus (Lohr, 1999). On tente ainsi d'améliorer la représentativité par taille sachant qu'il existe une certaine corrélation entre le nombre d'employés et le revenu.

Une autre limite du plan d'échantillonnage est reliée à l'estimation des ventes par Internet. Comme mentionné dans la section précédente, il s'agit d'un événement rare et nous avons peu d'information auxiliaire disponible sur la base de sondage pour cibler les entreprises faisant des ventes par Internet. Les estimations obtenues ne sont pas très précises pour cette variable. Au niveau industriel SCIAN2, représentant 22 secteurs au Canada, plus de la moitié des estimations ont un CV supérieur à 25%. En 2001, près du tiers des secteurs avaient un CV supérieur à 50%. Par contre, une légère amélioration est observée en 2002 et 2003, les CV de plus de 50% représentant seulement un dixième des 22 estimations produites.

Certaines mesures ont été prises permettant d'améliorer légèrement l'estimation des ventes par Internet au fil des années. Par exemple, nous sur-échantillonnons dans les secteurs plus propices au commerce électronique tels la fabrication, le commerce de gros, le commerce de détail, le transport et l'entreposage. De plus, nous forçons dans l'échantillon les plus grands contributeurs connus dans le domaine des ventes par Internet ainsi que les plus grandes entreprises au Canada. On fait l'hypothèse que les plus grandes entreprises sont plus propices au commerce électronique et à des ventes par Internet non négligeables. Finalement, des règles de vérification avec données historiques à la collecte permettent de corriger ou de valider les valeurs du répondant en faisant les suivis appropriés.

5. MÉTHODOLOGIE PROPOSÉE

Afin de mieux répondre aux besoins de l'enquête et des utilisateurs, nous proposons un nouveau plan d'échantillonnage. Ce plan devrait permettre d'améliorer deux aspects de l'enquête: une meilleure comparabilité dans le temps et une meilleure représentativité par taille.

Tout d'abord, on voulait stabiliser la population cible dans le temps c'est-à-dire la rendre moins sensible aux fluctuations et/ou corrections du revenu tiré du registre des entreprises. Pour ce faire, nous avons proposé de travailler avec une borne d'exclusion fixe au lieu d'une couverture de revenu de 95%. Nous avons considéré différentes options, en conservant autant que possible une population comparable à celle déjà utilisée et en utilisant différentes bornes de revenu par industrie. L'option s'est arrêtée sur deux bornes de revenu possibles : 250 000$ et 100 000$. Pour chaque secteur industriel, on choisit une des deux bornes d'exclusion permettant de couvrir 95% du revenu de la population totale. Par exemple, pour le secteur de la fabrication, en excluant de la population cible toutes les unités de moins de 250 000$ en revenus, on conserve facilement une couverture d'au moins 95% du revenu total du secteur. Par contre, dans le secteur de la restauration et de l'hébergement, un revenu de 100 000$ et plus est nécessaire afin de couvrir le 95%. Afin de stabiliser la population cible, il a été proposé de garder ces bornes pour les deux ou trois prochaines années. Cette méthode permet de conserver une taille de population similaire à l'année précédente (environ 650 000 entreprises).

Un autre besoin pour fin d'analyse est l'estimation par taille d'entreprises : petites, moyennes et grandes. On veut évaluer l'utilisation de la technologie pour les petites, moyennes et grandes entreprises et également leur évolution dans le temps. Nous voulions donc modifier l'échantillonnage afin de mieux tenir compte des définitions de petites, moyennes et grandes entreprises utilisées par les analystes de l'enquête.

Tout d'abord, le registre des entreprises de Statistique Canada contient la variable "nombre d'employés" disponible pour toutes les entreprises. Nous avons vérifié la corrélation de cette variable avec le nombre d'employés mesuré dans l'enquête. La corrélation étant très bonne à l'intérieur des classes d'intérêt, nous avons utilisé le nombre d'employés sur le registre des entreprises dans l'élaboration du plan d'échantillonnage. Nous avons testé plusieurs scénarios sous la contrainte de forcer dans l'échantillon les grandes entreprises, plus susceptibles de faire des ventes non négligeables par Internet. Le scénario retenu est formé de 4 strates à l'intérieur des domaines d'intérêt :

Strate 1: entreprises avec moins de 20 employés
Strate 2: entreprises de 20 à 99 employés
Strate 3: entreprises de 100 à 499 employés
Strate 4: entreprises de 500 employés et plus

Les strates 1 à 3 seront à tirage partiel et la strate 4 à tirage complet. Basé sur les estimations historiques, la strate 4 permettrait à elle seule de représenter 45% des ventes par Internet. Pour les strates 1 à 3, nous avons réparti la taille d'échantillon en utilisant la répartition de Neyman et la variable de revenu. L'avantage d'utiliser le revenu est d'obtenir un plus grand échantillon dans les strates moins homogènes selon le revenu, permettant ainsi de meilleures estimations économiquement pondérées.

Lors de la répartition de l'échantillon, nous avons sélectionné plus d'unités dans les secteurs propices aux ventes par Internet. Après avoir testé plusieurs scénarios, il a été décidé de consacrer la moitié de l'échantillon aux secteurs plus propices au commerce électronique même si ces secteurs ne représentent que le tiers de la population globale.

Finalement, comme dans l'ancien plan d'échantillonnage, nous avons tiré un échantillon sous la contrainte de maximiser le chevauchement avec l'échantillon de l'année précédente. La méthode de Kish et Scott (1971) a également été utilisée permettant un chevauchement de près de 60% avec l'échantillon de l'année précédente.

6. RÉÉVALUATION DES BESOINS VS LA MÉTHODOLOGIE PROPOSÉE

Nous résumons dans cette section comment les besoins de l'enquête définis dans la section 3 devraient être satisfaits suite à l'application de la méthodologie proposée.

Premièrement, en ce qui concerne la comparaison des estimations dans le temps, il est évident que le nouveau plan avec une population beaucoup plus stable et des bornes de stratification fixes pour toutes les industries va permettre une meilleure analyse des variables qualitatives mesurées.

Il en va de même pour l'analyse des estimations par taille d'entreprise définie selon le nombre d'employés, la stratification ayant été basée sur cette variable avec les mêmes définitions de taille.

En ce qui concerne l'estimation des ventes totales par Internet, nous ne pensons pas avoir un plan qui améliore cette variable. Nous utilisons la même base (le registre des entreprises) et nous avons donc les mêmes limites qu'auparavant (pas d'information sur les ventes par Internet, à l'exception du secteur industriel et du revenu). Par contre, nous ne prévoyons pas une diminution de la qualité et nous envisageons d'autres avenues dans le futur pour améliorer cette variable (voir la prochaine section).

Finalement, même si ce n'est plus un plan d'échantillonnage optimum pour les estimations économiquement pondérées, nous croyons que le nouveau plan est assez robuste. Les plus grandes unités sont dans des strates auto-représentatives et la taille de l'échantillon a été répartie par strate en tenant compte de la variabilité des revenus d'entreprises dans les strates (répartition de Neyman).

Pour la prochaine année, il a été décidé d'utiliser un échantillon sous le nouveau plan en parallèle avec un échantillon sous l'ancien plan. Ces échantillons en parallèle permettront d'évaluer l'impact du nouveau plan d'échantillonnage sur les estimations et d'ajuster les données historiques au besoin.

7. DÉFIS À VENIR

Un des plus grands défis à venir sera l'amélioration de la qualité de l'estimation des ventes totales par Internet. Comme mentionné dans l'article, il s'agit d'un événement rare et nous avons peu d'information auxiliaire pour cibler les entreprises faisant des ventes par Internet. Les estimations obtenues ne sont pas très précises pour cette variable. De plus, au cours des années, suite à des comparaisons historiques et des suivis, nous avons remarqué que cette variable n'était pas toujours comprise par les répondants. Dans les deux cas, certaines mesures ont été prises telles l'inclusion dans l'échantillon d'entreprises faisant des ventes considérables par Internet, un plus grand échantillon dans les secteurs plus propices au commerce électronique, des suivis auprès des répondants et finalement des clarifications sur le questionnaire.

Un des défis pour cette variable est l'utilisation d'une nouvelle base ou d'une base complémentaire contenant de l'information sur des entreprises faisant des ventes par Internet. Nous nous sommes demandés si l'utilisation d'Internet et d'un outil de recherche permettraient d'identifier de telles entreprises. Chose certaine, si ces entreprises font des ventes par Internet, elles se retrouvent forcément sur ce moyen de communication. La question est de savoir si l'extraction d'une base de données à partir d'Internet serait de bonne qualité et en mesure de couvrir notre population d'intérêt. Les techniques utilisées pour la fouille de données (Data mining), plus spécifiquement l'analyse de données textuelles (Text mining), pourraient être envisagées (Chauchat et Morin, 2004). L'analyse de données textuelles, en utilisant l'analyse des correspondances, permet de trouver des associations entre les mots ou encore des associations entre des documents. Dans notre cas, il s'agirait de définir un vocabulaire commun de sites d'entreprises faisant des ventes par Internet (par exemple, en utilisant les entreprises de l'enquête ayant rapportées des ventes par Internet) et de trouver par l'analyse des correspondances un vocabulaire commun à ces sites. Par la suite, on utilise ces mots communs comme mots de recherche dans Internet.

Un autre défi est d'évaluer la compréhension des répondants en ce qui concerne cette variable : comprennent-ils et rapportent-ils ce qu'on veut mesurer exactement (inclusions/exclusions)? Ont-ils l'information disponible? Il serait

tout indiqué pour notre enquête de refaire des groupes de discussions auprès d'entreprises même si, à l'origine de l'enquête, des groupes de discussion avait été mis en place pour tester le questionnaire. On doit se rappeler que depuis 1999, l'expertise s'est développée et de nombreux changements ont été apportés à l'enquête (changement de l'unité échantillonnale, changement du questionnaire, rétroaction obtenue suite aux suivis auprès des répondants,…). Les discussions pourraient porter sur le commerce électronique, mais également sur le fonctionnement en général de l'utilisation de la technologie par les entreprises (par exemples, comment mettent-elles en place un site Web et comment le maintiennent-elles? À quel niveau de l'entreprise les décisions sont-elles prises sur l'utilisation ou non de certaines technologies?).

8. CONCLUSION

Une enquête économique est constamment en évolution : les besoins changent, s'accroissent; l'expertise se développe, les sources d'information disponibles se multiplient et les entreprises évoluent. Une enquête est donc sujette aux changements au fil des années afin de bien et de mieux répondre aux besoins.

L'enquête sur le commerce électronique et la technologie est une enquête assez récente portant sur un domaine en constante évolution. L'information doit être disponible rapidement sinon elle n'est plus à jour, les besoins en données plus détaillées s'accroissent au fil des années et la comparaison historique est primordiale afin de mesurer adéquatement l'évolution de l'utilisation de la technologie dans le temps. Nous croyons qu'un nouveau plan d'échantillonnage était nécessaire pour mieux répondre à ces besoins. Le plan choisi permettra très certainement une meilleure comparaison dans le temps ainsi qu'une meilleure analyse par taille d'entreprises.

Un des défis qui demeure est l'amélioration de la variable des ventes totales par Internet, une des variables les plus importantes de l'enquête. Notre prochaine étape sera de mieux mesurer cet événement rare. Nous avons un outil disponible, l'Internet, qui contient une partie de l'information. Il n'en tient qu'à nous d'en évaluer le potentiel!

RÉFÉRENCES

Chauchat, J.-H. et Morin, A. (2004), "Data Mining et Text Mining", atelier présenté au *XXIe Symposium International sur les Méthodes innovatrices pour enquêter auprès des populations difficiles à joindre*, Gatineau, Canada.

Cuthill, I. (1997), "Le registre des entreprises de Statistique Canada", rapport non publié, Ottawa, Canada, Statistique Canada.

Estevao, V. M. (2005), "GES V4.3 User Guide", rapport non publié, Ottawa, Canada, Statistique Canada.

Kish, L. et Scott, A. (1971), "Retaining Units after Changing Strata and Probabilities", *Journal of the American Statistical Association*, 66, pp. 461-470

Lavallée P. et Hidiroglou M. (1988), "Sur la stratification de population asymétriques", *Techniques d'enquête*, 14, pp. 35-45.

Lohr, S. L. (1999), "Non-response", *Sampling Design and Analysis*, Press, pp. 255-282.

Nobrega, K. (1998), "Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method", *1998 Proceedings of the Survey Methods Section*, Statistical Society of Canada, pp. 137-141.