Methodology of the Survey of Electronic Commerce and Technology 2000 (SECT)

1. Introduction

The Survey of Electronic Commerce and Technology 2000 (SECT) is an annual survey. It collects information on communication and technology such as the use of computers, Internet and web sites, as well as the use of Internet to do electronic commerce. The SECT uses the sample from a well-known survey at Statistics Canada, the Capital Expenditures Survey (CES).

The collection began in November 2000 and data for the reference year 2000 was published in April 2001.

Note that the following sections, up to the section 3.5.6, describe the methodology used by the CES. Then, the methodology is specific to the SECT. The mainly difference between the two surveys are the units of reference. The CES collects information at the establishment level while the SECT collects information at the enterprise level.

Coverage by Industrial Sector

The sample used for this survey covers most industrial sectors. These are described using the North American Industrial Classification System (NAICS). In the sampling design of the CES, some sectors are left out for different reasons. These exceptions are now SECT exceptions as well. Here is the list:

A) **Sector 11 Sub-sector 111, 112 and 114** (Crop and Animal Production Industries, Fishing, hunting and Trapping industries)
B) **Sector 23** (Construction Industry)
C) **Sector 91 Sub-sector 913** (Local Governments)
D) **Sector 21 Canadian Industry 213119** (Other support activities for mining)
E) **Sector 55 Canadian Industry 551114** (Head office),
F) **Sector 81 Sub-sector 814** (Private households).

2. Survey Frame

The frame consists primarily of the Business Register (**BR**) developed by Statistics Canada. Business Register Division (**BRD**) is responsible for the maintenance and updating of the register. The register is used by a large number of surveys that in turn provide it with feedback to ensure that the latest changes in the business world are incorporated into the BR as quickly as possible.

The BR contains the units required to establish our final survey frame. They are arranged hierarchically as

Méthodologie de l'enquête sur le commerce électronique et la technologie 2000 (ECET)

1. Introduction

L'enquête sur le commerce électronique et la technologie 2000 (ECET) est une enquête annuelle. Elle collecte de l'information sur les communications et la technologie tels l'utilisation de l'ordinateur, l'Internet et les sites Web, ainsi que l'utilisation de l'Internet à des fins de commerce électronique. L'ECET utilise l'échantillon d'une enquête bien connue de Statistique Canada, l'enquête sur les dépenses en immobilisations (EDI).

Les envois postaux ont débutés en novembre 2000 et des chiffres pour l'année 2000 ont pu être publiés dès avril 2001.

À noter que les section suivantes, jusqu'à la section 3.5.6, décrivent la méthodologie utilisée par l'EDI. Par la suite, la méthodologie est spécifique à l'ECET. Notons finalement que les deux enquêtes diffèrent principalement par l'unité de référence. L'EDI recueille l'information au niveau de l'établissement alors que l'ECET recueille l'information au niveau de l'entreprise.

Couverture par secteur industriel

L'échantillon utilisé pour cette enquête couvre à peu près tous les secteurs industriels. Ceux-ci sont décrits en utilisant la convention connue sous le Système de classification industriel de l'Amérique du Nord (SCIAN). Puisque l'échantillon de l'ECET est en fait celui de l'EDI, il doit subir certaines contraintes dont notamment, de ne pas couvrir exactement tous les secteurs industriels. Voici donc les exceptions :

A) **Secteur 11 sous-secteurs 111, 112 et 114** (Industrie de la production animale et agricole, Industrie de la pêche, de la chasse et du piégeage),
B) **Secteur 23** (Industrie de la construction),
C) **Secteur 91 sous-secteur 913** (Administrations locales),
D) **Secteur 21 industrie canadienne 213119** (Services reliés aux mines),
E) **Secteur 55 industrie canadienne 551114** (Bureaux-Chefs),
F) **Secteur 81 sous-secteur 814** (Ménages privés).

2. Base de sondage

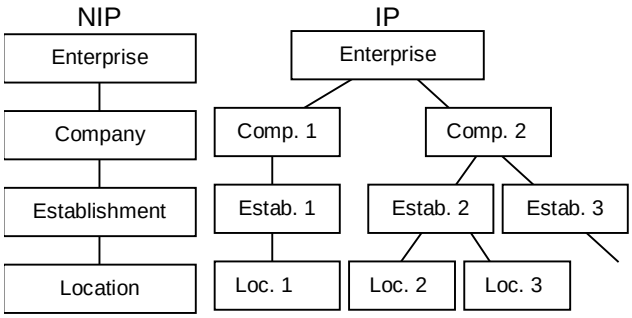
La base de sondage est principalement formée du Registre des entreprises (**RE**) développé à Statistique Canada. La Division du registre des entreprises (**DRE**) est chargée d'en faire l'entretien et la mise à jour. Le registre est utilisé par un grand nombre d'enquêtes qui ne manquent pas de lui retourner de la rétroaction pour s'assurer que les plus récents changements dans le monde des entreprises soient incorporés au RE dans les plus brefs délais.

On retrouve sur le RE les unités nécessaires à l'établissement de la base de sondage finale. La hiérarchie s'y lit comme suit : Entreprise - Compagnie - Établissement - Emplacement. Une

follows: Enterprise – Company - Establishment - Location. An enterprise may comprise several companies, of which each may have several establishments that in turn may operate in several locations. This so-called “statistical” structure is in fact a model of the operational structure that is described by the enterprise itself. Based on the information available for each level of the operational structure, we define the corresponding statistical structure. For example, to be considered an establishment, an establishment should be able to supply the BR with the wages and rates of pay, income and major inputs in the operational process. To be considered an enterprise, an enterprise should be able, for example, to supply the BR with its organisational structure and its assets.

There are two kinds of units on the BR: the units that are part of the non-integrated portion (**NIP**) and the units that are part of the integrated portion (**IP**). The units from the NIP are small and medium units with a statistical structure that is linear: an enterprise is related to a single company, a single establishment and a single location. The units from the IP are large units with a statistical structure that may be linear but usually is more complex. Diagram 2 illustrates both structures.

Diagram 2: Statistical Structures



The sampling unit selected for the Capital Expenditure Survey is the establishment, which best corresponds to the gathering and disclosure of investment data. For more details on the BR, refer to Cuthill (1996).

Note that the sample used for the CES is the one selected in November of the previous year, since this is the last sample selected and ready for the SECT collection, which begins in November. Because of the dynamic nature of businesses, we can be certain that new establishments started up between the time the sample was selected and the time the collection began. To obtain a better coverage for the desired reference year, we added units to the original sample. These units are found by using a newer version of the BR, from newspapers, company reports or lists of available administrative data. These units are sampled with certainty.

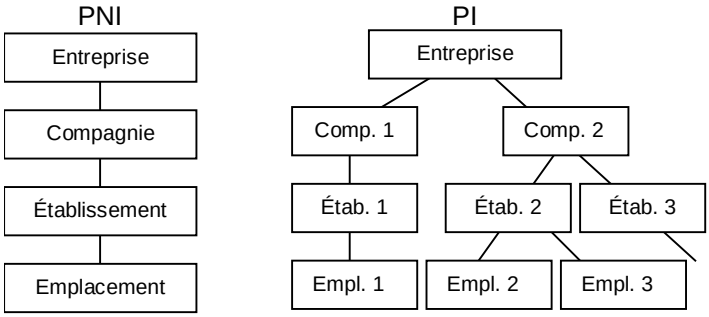
Grouping

Before sampling begins, all units from the private sector not in the mining and manufacturing industries are grouped together using the following method. All

entreprise peut avoir plusieurs compagnies qui peuvent chacune avoir plusieurs établissements qui peuvent à leur tour avoir plusieurs emplacements. Cette structure dite «statistique» est en fait une modélisation de la structure opérationnelle décrite par l’entreprise elle-même. Selon l’information disponible pour chaque niveau de la structure opérationnelle, on définit le niveau statistique correspondant. Par exemple, pour être considéré comme un établissement, on devrait pouvoir fournir au RE les salaires et taux de rémunération, le revenu et les intrants principaux dans le processus d’exploitation. Pour être considéré comme une entreprise, on devrait par exemple pouvoir fournir au RE la structure organisationnelle et les actifs de l'entreprise.

Il existe deux type d'unités sur le RE: les unités formant la portion non-intégrée (**PNI**) et les unités formant la portion intégrée (**PI**). Les unités du PNI sont des unités de petites et moyennes tailles et dont la structure statistique est linéaire: une entreprise est reliée à une seule compagnie, à un seul établissement et à un seul emplacement. Les unités du PI sont des unités de grandes tailles et dont la structure statistique peut être linéaire mais est généralement plus complexe. Le schéma 2 illustre les deux structures.

Schéma 2: Structures statistiques



Dans le cadre de l’EDI, l’unité d’échantillonnage choisie est l’établissement, celle-ci correspondant le mieux au besoin de collecte et de divulgation des données d’investissements. Pour plus de détails concernant le RE, consultez Cuthill (1996).

Il est à noter que l’échantillon de l’EDI utilisé est celui de novembre de l’année précédente, puisqu’il s’agit du dernier échantillon sélectionné et prêt pour la collecte de l’ECET qui débute en novembre. Étant donné la nature dynamique des entreprises, il est certain que de nouveaux établissements ont été mis sur pied entre la période où l’échantillon est tiré et le début de la collecte. Afin d’obtenir une meilleure couverture pour l’année de référence voulue, on ajoute des unités à l’échantillon original. Ces unités sont trouvées, soit en utilisant une version plus récente du RE ou encore par la lecture de journaux, de rapports de compagnie ou encore grâce aux listes administratives disponibles. Les unités ajoutées sont échantillonnées avec certitude.

Regroupement

Avant de procéder à l’échantillonnage, toutes les unités dans le secteur privé qui ne faisaient pas partie des industries minières et manufacturières ont été regroupées selon la méthode suivante.

establishments operating in the same province, in the same six-digit-code industrial sector and under the same enterprise have been grouped together in a single super-establishment. The income of the super-establishment is the sum of all income for the establishments that comprise it, while the remaining information is taken from the head of the group, either the head office where possible, or the establishment with the highest income, where applicable. For the public sector, all the units are in the sample

Once the new universe is constructed with the new super-establishments, all units with income of less than a certain limit are eliminated from the frame. We consider these units with negligible investments. The exclusion of these units allows us to reduce the response burden of the respondents. However, the head offices and laboratories are kept in the population and are selected in the sample with certainty. This procedure is instituted to avoid excluding these units, which generate practically no income, but might account for substantial investment.

The limit that delineates the out-of-scope units is determined as a function of province and industry. It varies from $100,000 to $250,000 depending on the size of the units in the grouping. The limit is calculated in such a way that a maximum of 5% of the total revenue in the group becomes out-of-scope. When all groups have been assembled and the out-of-scope units have been eliminated, the universe is ready for stratification.

3. Sampling

The sampling is divided into the three traditional parts: stratification, allocation and selection. These are described in the following text.

Stratification

The sample must first be stratified by geographic location and industrial classification. The geographic division is based on the 13 provinces and territories, with no other refinement (no infra-provincial stratification). For the industrial stratification, the 1997 NAICS is used at the level required for estimation purposes. If, for example, for a certain industry, the most disaggregated level published corresponds to the 3-digit NAICS, this will be the stratification level. It should be noted that for the remainder of the section, the 6-digit NAICS will be abbreviated as NAICS-6, the 5-digit NAICS as NAICS-5, and so forth.

Table 1 shows, by industry, the level of the stratification.

Table 1
Stratification Levels

Industry Sector	NAICS Stratification Level
11 - Agriculture, Forestry, Fishing and Hunting	3
21 - Mining and Oil and Gas Extraction (NAICS-3 213)	3

Tous les établissements opérant dans la même province, dans le même secteur industriel codé à six chiffres et sous la même entreprise ont été regroupés en un seul super-établissement. Le revenu du super-établissement est la somme de tous les revenus des établissements qui le composent et le reste de l'information est tiré de la tête du regroupement, soit le bureau-chef si c'est possible ou, sinon, l'établissement avec le plus grand revenu. Pour le secteur public, toutes les unités font partie de l'échantillon.

Une fois le nouvel univers construit avec les nouveaux super-établissements, toutes les unités qui ont un revenu inférieur à une certaine limite sont éliminées de la base. On considère ces unités ayant des investissements négligeables. L'exclusion de ces petites unités permet de réduire le fardeau de réponse des répondants. Par contre, les bureaux-chefs ou les laboratoires sont gardés dans la population et sont choisies dans l'échantillon avec certitude. Cette procédure est mise en place pour éviter d'exclure ces unités qui ne génèrent pratiquement aucun revenu, mais qui pourraient être l'objet d'investissements substantiels.

La limite inférieure déterminant les unités dans le champ de l'enquête est construite en fonction de la province et du secteur industriel. Celle-ci varie de 100 000$ à 250 000$ dépendamment de la taille des unités qui composent l'industrie. En gros, la limite est calculée de telle sorte qu'un maximum de 5% du revenu total du secteur industriel devient hors champs. Lorsque tous les regroupements ont été effectués et que les unités hors champs ont été éliminées, l'univers est prêt à être stratifié.

3. Échantillonnage

L'échantillonnage se divise selon les trois parties traditionnelles: la stratification, la répartition et la sélection. Celles-ci sont décrites dans le texte qui suit.

Stratification

On doit tout d'abord stratifier selon le lieu géographique et la classification industrielle. La division géographique se fait selon les 13 provinces et territoires, sans autre raffinement (aucune stratification infra-provinciale). Pour ce qui est de la stratification industrielle, le SCIAN de 1997 est utilisé selon le niveau requis pour les estimations. Si par exemple, pour une certaine industrie, le niveau le plus désagrégé publié correspond au SCIAN à 3 chiffres, ce sera le niveau de stratification. Notons que pour le reste de la section, le SCIAN à 6 chiffres sera abrégé par SCIAN-6, le SCIAN à 5 chiffres par SCIAN-5, etc.

Le tableau 1 indique, par industrie, quels sont les niveaux de stratification.

Tableau 1
Niveaux de stratification

Secteur industriel	Niveau de stratification SCIAN
11 - Agriculture, foresterie, pêche et chasse	3
21 - Extraction minière, de pétrole et gaz (SCIAN-3 213)	3
21 - Extraction minière, de pétrole et gaz (SCIAN -5 21231 et 21232)	5

21 - Mining and Oil and Gas Extraction (NAICS-5 21231 and 21232)	5
21 - Mining and Oil and Gas Extraction (Other NAICS)	6
22 - Utilities	4
31-33 Manufacturing (NAICS-3 316 and 323)	3
31-33 Manufacturing (NAICS-4 3121 and NAICS-3 322, 324 and 326)	5
31-33 Manufacturing (Other NAICS)	4
41 - Wholesale Trade	3
44-45 - Retail Trade	3
48-49 – Transportation and Warehousing	3
51 - Information and Cultural Industries	3
52 - Finance and Insurance	3
53 - Real Estate and Rental and Leasing	4
54 - Professional, Scientific and Technical Services	4
55 – Management of Companies and Enterprises	2
56 – Administration and Support, Waste Management and Remediation Services	3
61 - Education Services	4
62 - Health Care and Social Assistance	3
71 - Arts, Entertainment and Recreation	3
72 – Accommodations and Food Services	3
81 - Other Services	3
91 - Public Administration	3

Allocation

The allocation consists in determining the sample size required for an expected level of precision (CV) (see section 3.5.11 for more information on CV) and in allocating the sample size in the strata.

Once the initial stratification has been introduced, we compute the coefficient of variation (**CV**) to be targeted using the revenue variable to reach the CV set for the most disaggregated publication level. An example helps to better define the situation.

Assume that we want to publish estimates for sector 72 (Accommodations and Food Services), which corresponds to NAICS-3 at the Canada level and the whole industry at the Province / Territory level. We then construct Table 2, in which the number of provinces has been reduced to 3 and the number of NAICS-3 for the industry as a whole is 2, specifically the sub-sectors (**SS**) 721 and 722.

Table 2
Cross Publication for Sector 72

	Prov1	Prov2	Prov3	CV
SS721				15%
SS722				15%
CV	15%	15%	15%	

The initial stratification corresponds to each cell in table 2 and the marginals correspond to the estimates we wish to publish. If, for example, we wish to publish esti-

21 - Extraction minière, de pétrole et gaz (autres SCIAN)	6
22 - Services publics	4
31-33 Fabrication (SCIAN -3 316 et 323)	3
31-33 Fabrication (SCIAN -4 3121 et SCIAN -3 322, 324 et 326)	5
31-33 Fabrication (autres SCIAN)	4
41 - Commerce de gros	3
44-45 - Commerce de détail	3
48-49 - Transport et entreposage	3
51 - Information et culture	3
52 - Finance et assurances	3
53 - Services immobiliers, de location et de location à bail	4
54 - Services professionnels, scientifiques et techniques	4
55 - Gestion de sociétés et d'entreprises	2
56 - Services administratifs, de soutien, de gestion des déchets et d'assainissement	3
61 - Services d'enseignement	4
62 - Soins de santé et assistance sociale	3
71 - Arts, spectacles et loisirs	3
72 - Hébergement et services de restauration	3
81 - Autres services	3
91 - Administrations publiques	3

Répartition

La répartition consiste à déterminer la taille d'échantillon requise pour un niveau de précision voulue (CV) (voir la section 3.5.11 pour plus d'information sur les CV) et de répartir la taille de l'échantillon dans les strates.

La première étape consiste, à partir de la stratification initiale définie précédemment, à calculer le coefficient de variation (**CV**) à viser en utilisant la variable revenu de façon à atteindre le CV fixé pour le niveau de publication le plus désagrégé. Un exemple aide à mieux comprendre la situation.

Supposons qu'on veuille publier des estimations pour le secteur industriel 72 (Hébergement et services de restauration) pour lequel on publie au niveau SCIAN-3 pour le Canada et au niveau de l'industrie complète par province / territoire. On construit alors le tableau 2, où le nombre de provinces a été simplifié à 3 et le nombre de sous-secteurs (**SS**) SCIAN-3 pour l'industrie au complet est 2 (721 et 722).

Tableau 2
Croisements de publication pour le secteur 72

	Prov1	Prov2	Prov3	CV
SS721				15%
SS722				15%
CV	15%	15%	15%	

La stratification initiale correspond à chacune des cellules du tableau 2 et les marginales correspondent aux estimations qu'on désire publier. Si on désire, par exemple, publier des estimations avec un CV cible de 15%, on doit d'abord calculer le CV à viser

mates with a target CV of 15%, we must first compute the CV to be targeted for each cell, so that the marginal CVs are met.

Before we can compute the CV required at the cell level to reach the CV set for the marginals, we must adjust the marginal CVs in order to ensure that the total variance at the Canada level is the same when calculated at the industry level or when calculated at the provincial level. In fact, we cannot obtain 15% CVs in both directions, because when we set the variance in one direction to obtain the targeted CV, we automatically set the variance (thus the CV) for the other direction and we are "subject to" the resulting CV. With the knowledge that the CVs in both directions cannot be simultaneously equal to the targeted CV (unless by chance), we have chosen to minimize the distance from the marginal CVs to the target CV. In one direction, we then obtain a resulting CV greater than the target CV and in the other, a CV less than this same CV. This is done by minimizing the distance between the resulting CVs and the target CV under the constraint that the variances must be the same in both directions. In mathematical terms:

$$\text{Minimize } (CV^c - CV^A)^2 + (CV^c - CV^B)^2$$
$$\text{under the constraint } V^A = V^B$$

where CV^A and CV^B represent the CVs attainable in both directions, CV^c represents the target CV and V^A and V^B represents the variances in both directions.

Let us call the resulting CV the new target CV. In the preceding example, we could end up with new target CVs as in Table 3.

Table 3

New target CVs(closest to the targeted CV)

	Prov1	Prov2	Prov3	CV
SS721				11%
SS722				11%
CV	18%	18%	18%	

To reach the new target CV, we must compute what the targeted CVs should be for each of the initial strata by using a raking ratio algorithm as described in Latouche (1988).

Using the letters A and B again to designate the two directions (A the geographic direction and B the industrial direction, for example), we recompute the cell CVs until the combination of the CVs on the same line or in the same column is close enough to the target CV for the corresponding marginal.

pour chacune des cellules de telle sorte que les CV des marginales soient respectés.

Avant de pouvoir calculer le CV nécessaire au niveau des cellules pour atteindre le CV fixé au niveau des marginales, on doit d'abord ajuster les CV marginaux de sorte que la variance totale au niveau Canada soit la même, qu'elle soit calculée au niveau industriel ou qu'elle soit calculée au niveau provincial. En effet, on ne peut obtenir des CV de 15% dans les deux directions, car lorsque l'on fixe la variance dans une direction pour obtenir le CV visé, on fixe automatiquement la variance (donc le CV) pour l'autre direction et on «subit» le CV résultant. Sachant que les CV des deux directions ne peuvent être simultanément égaux au CV visé (à moins d'un hasard), nous avons choisi de minimiser la distance des CV des marginales au CV cible. On obtient donc, dans une direction, un CV résultant supérieur au CV cible et dans l'autre, un CV inférieur à ce même CV. Ceci est fait en minimisant la distance entre les CV résultants et le CV cible sous la contrainte d'avoir des variances égales dans les deux directions. D'une façon mathématique:

$$\text{Minimiser } (CV^C - CV^A)^2 + (CV^C - CV^B)^2$$
$$\text{sous la contrainte } V^A = V^B$$

où CV^A et CV^B représentent les CV atteignables dans les deux directions, CV^C représente le CV cible et V^A et V^B représentent les variances dans les deux directions.

Appelons le CV résultant le nouveau CV cible. Dans l'exemple précédent, on pourrait se retrouver avec de nouveaux CV cibles comme dans le tableau 3.

Tableau 3

Nouveaux CV cibles (les plus près du CV visé)

	Prov1	Prov2	Prov3	CV
SS721				11%
SS722				11%
CV	18%	18%	18%	

Pour atteindre le nouveau CV cible, on doit calculer ce que devraient être les CV visés pour chacune des strates initiales en utilisant l'algorithme itératif du quotient tel que décrit dans Latouche (1988).

En utilisant à nouveau les lettres A et B pour désigner les deux directions (A la direction géographique et B la direction industrielle par exemple), on recalcule les CV des cellules jusqu'à ce que la combinaison des CV sur une même ligne ou une même colonne soit assez près du CV cible de la marginale correspondante.

$$CV_r^B(\hat{Y}_{ij}) = CV_{(r-1)}^A(\hat{Y}_{ij}) * \frac{CV(\hat{Y}_{.j})\hat{Y}_{.j}}{\sqrt{\sum_j (CV_{(r-1)}^A(\hat{Y}_{ij}))^2 \hat{Y}_{ij}^2}}$$

$$CV_r^B(\hat{Y}_{ij}) = CV_{(r-1)}^A(\hat{Y}_{ij}) * \frac{CV(\hat{Y}_{.j})\hat{Y}_{.j}}{\sqrt{\sum_j (CV_{(r-1)}^A(\hat{Y}_{ij}))^2 \hat{Y}_{ij}^2}}$$

$$CV_r^A(\hat{Y}_{ij}) = CV_{(r-1)}^B(\hat{Y}_{ij}) * \frac{CV(\hat{Y}_{i.})\hat{Y}_{i.}}{\sqrt{\sum_j (CV_{(r-1)}^B(\hat{Y}_{ij}))^2 \hat{Y}_{ij}^2}}$$

where
- r denotes the current iteration,
- r-1 denotes the preceding iteration,
- i. denotes the marginal in direction A,
- .j denotes the marginal in direction B,
- ij denotes a crossover of directions A and B and
- Y corresponds to the total for the income variable for a given group.

The algorithm stops when the convergence criterion (0.1%) is met or after a maximum of 10 iterations. It should be noted here that the algorithm converges very quickly and is almost certain to reach the targeted CV for the marginals. Table 4 illustrates the result of the iterative procedure.

Table 4
Cell CVs after Iterations

	Prov1	Prov2	Prov3	CV
SS721	20%	23%	24%	11%
SS722	17%	20%	21%	11%
CV	18%	18%	18%	

Now that the CV is set for each of the initial strata (these correspond to the cells in the preceding table), we can stratify them into two major strata: large, in which the sample is conducted with certainty, and small, in which the sampling is conducted under a probability scheme so the new target CV can be attained. The preferred method for splitting cells in two is that advanced by Hidiroglou (1986) which has the merit of minimizing the sampling size while attaining the target CV. The technique is simple: start with the equation that gives the CV for the initial stratum

$$CV(\hat{Y})^2 = \frac{\frac{(N-t)*(N-n(t))}{(n(t)-t)} S_{(N-t)}^2}{\hat{Y}^2}$$

where
- N denotes the population size,
- n(t) denotes the total number of units to be sampled,
- t denotes the total number of units in the take-all stratum,
- S^2 (n-t) denotes the variance in the take-some stratum and
- Y corresponds to the total of the income

$$CV_r^A(\hat{Y}_{ij}) = CV_{(r-1)}^B(\hat{Y}_{ij}) * \frac{CV(\hat{Y}_{i.})\hat{Y}_{i.}}{\sqrt{\sum_j (CV_{(r-1)}^B(\hat{Y}_{ij}))^2 \hat{Y}_{ij}^2}}$$

où
- r désigne l'itération courante,
- r-1 désigne l'itération précédente,
- i. désigne la marginale dans la direction A,
- .j désigne la marginale dans la direction B,
- ij désigne un croisement des directions A et B et
- Y correspond au total de la variable revenu pour un groupement donné.

L'algorithme s'arrête lorsque le critère de convergence (0,1%) est rencontré ou après un maximum de 10 itérations. Notons ici que l'algorithme converge très rapidement et on atteint presque à coup sûr le CV visé au niveau des marginales. Le tableau 4 illustre le résultat du procédé itératif.

Tableau 4
CV des cellules après itérations

	Prov1	Prov2	Prov3	CV
SS721	20%	23%	24%	11%
SS722	17%	20%	21%	11%
CV	18%	18%	18%	

Maintenant que le CV est fixé pour chacune des strates initiales (celles-ci correspondent aux cellules du tableau précédent), on peut les stratifier en deux strates de taille: grande taille où l'échantillonnage se fait avec certitude et petite taille où l'échantillonnage se fait selon une probabilité de sélection permettant d'atteindre le nouveau CV cible. La méthode préconisée pour séparer les cellules en deux est celle d'Hidiroglou (1986) qui a le mérite de minimiser la taille échantillonnale tout en atteignant le CV cible. La technique est simple: on part de l'équation qui donne le CV de la strate initiale

$$CV(\hat{Y})^2 = \frac{\frac{(N-t)*(N-n(t))}{(n(t)-t)} S_{(N-t)}^2}{\hat{Y}^2}$$

où
- N désigne la taille de la population,
- n(t) désigne le nombre total d'unités à échantillonner,
- t désigne le nombre total d'unités dans la strate à tirage complet,
- S^2(n-t) désigne la variance dans la strate à tirage partiel et
- Y correspond au total de la variable revenu pour la strate.

On peut la réécrire de façon à isoler n(t), le nombre total d'unités à échantillonner en fonction de t, le nombre d'unités échantillonnées avec certitude:

variable for the stratum.

It can be rewritten to isolate n(t), the total number of units to be sampled based on t, the number of units sampled with certainty:

$$n(t) = t + \frac{(N-t)^2 S^2_{(N-t)}}{CV^2 \hat{Y}^2 + (N-t) S^2_{(N-t)}}$$

We then must clearly understand the function to find its minimum point. This can be attained through an iterative process that computes the following two parameters after converging: the dividing value separating the initial stratum into two final strata as well as the sample size for each of the strata. There will be t units in the take-all stratum and n(t) - t units to be taken in the take-some stratum. This process will have taken the minimum number of units to attain the target CV set.

It is highly likely that we will not obtain the precise target CV for the cells. The CV reached is usually close, but for some cells may be as much as 2% below the target CV. The effect of this is a slight change in the CVs targeted for the marginals. Table 5 reproduces the results from Table 4 following application of Hidiroglou's algorithm.

Table 5
Final Cell CVs after Iterations

	Prov1	Prov2	Prov3	CV
SS721	20.1%	22.8%	24%	10.8%
SS722	17.2%	21.5%	20.4%	11.7%
CV	18.1%	18.9%	17.8%	

Once this step is complete, we can then proceed with the actual selection of the sample.

Selection

For the take-some strata, selection is based on a simple random process. A minimal fraction of 1% is imposed and a minimum of 3 units is required where possible. In the take-all strata, all units are sampled with certainty. This selection method forces no unit into the sample and takes no account of the preceding sample.

Up to now, the strategy described for the SECT is exactly the same as the one in CES. The rest of the document is specific to the SECT and describes the collection, imputation and estimation methods, and begins with the specificity of the SECT.

$$n(t) = t + \frac{(N-t)^2 S^2_{(N-t)}}{CV^2 \hat{Y}^2 + (N-t) S^2_{(N-t)}}$$

Il s'agit alors de bien comprendre la fonction pour trouver son point minimum. Celui-ci peut être atteint selon un processus itératif qui calcule, après avoir convergé, les deux paramètres suivants: la borne qui sépare la strate initiale en deux strates finales ainsi que la taille échantillonnale de chacune des strates. On aura t unités dans la strate à tirage complet et n(t) - t unités à tirer dans la strate à tirage partiel. On aura ainsi tiré le nombre minimal d'unités pour atteindre le CV cible fixé.

Il est fort probable qu'on n'obtienne pas exactement le CV cible au niveau des cellules. Le CV atteint est habituellement près, mais peut être pour certaines cellules jusqu'à 2% au-dessus du CV cible. Ceci a pour effet de changer légèrement les CV visés au niveau des marginales. Le tableau 5 reprend les résultats du tableau 4 après l'application de l'algorithme d'Hidiroglou.

Tableau 5
CV final des cellules après itérations

	Prov1	Prov2	Prov3	CV
SS721	20.1%	22.8%	24%	10.8%
SS722	17.2%	21.5%	20.4%	11.7%
CV	18.1%	18.9%	17.8%	

Lorsque cette étape est complétée, on peut alors procéder à la sélection proprement dite de l'échantillon.

Sélection

Pour les strates à tirage partiel, la sélection se fait selon un processus aléatoire simple. Une fraction minimale de 1% est imposée et un minimum de 3 unités est exigé là où c'est possible. Dans les strates à tirage complet, toutes les unités sont échantillonnées avec certitude. Cette méthode de sélection ne force aucune unité dans l'échantillon et ne tient nullement compte de l'échantillon précédent.

Jusqu'à présent, la procédure suivie pour l'ECET est exactement la même que dans le cas de l'EDI. Le reste du texte est maintenant spécifique à l'ECET et décrit les procédures de collecte, de suivi, d'imputation et d'estimation en débutant par le caractère spécifique de l'ECET.

4. Spécificité de l'ECET

Contrairement à l'EDI qui recueille l'information au niveau de l'établissement et qui désire obtenir des estimations à ce niveau, l'ECET recueille l'information au niveau de l'entreprise et désire

4. Specificity of the SECT

Contrary to the CES which collects information at the establishment level and desires to get estimates at that level, the SECT collects information at the enterprise level and wants to get estimates at the enterprise level. For the CES, the establishment is chosen since it is the level for which we can get investment information. For the SECT, the management of electronic services and telectronic commerce are usually maintained at the enterprise level.

Since the CES sampling design is at the establishment level, we used the weight share method (Lavallée, 1995) to produce estimates at the enterprise level for the SECT. The method is described in the section 3.5.11 – Estimation.

The data for the SECT was collected at the enterprise level. To do that, we had to derive the enterprises linked to the sampled establishments. Usually, for each establishment on the BR, we can find an associated enterprise number. For the other cases, establishments were grouped in order to help the respondent to answer and also to get a group that is similar to the definition of the enterprise.

Also, for the needs of collection and imputation, we calculated an auxiliary variable at the enterprise level by using information on the BR or summing the information available at the establishment level. The auxiliary variable is the revenue for units in the private sector and the number of employees for the public sector. However, for certain units, such as head offices and units belonging to the public sector, we had to use other sources to obtain the auxiliary information. We used internal sources of Statistics Canada, namely data from the Survey of Employment, Payroll and Hours and data from the Public Institution Division. Finally, mean imputation within homogeneous groups allowed the creation of an auxiliary variable for the unsolved cases.

5. Collection and Data Editing

The sample originally selected for CES had already been used, and it was useful in determining that some establishments were out of business, amalgamated or simply duplicates. For the rest of the units, a questionnaire was mailed to the enterprise level and respondents were encouraged to complete and return it.

At data collection, some edits are applied to each questionnaire such as rules of consistency. For example, some rules verify that if some fields are coded, all related fields are also coded. For more details on the edit rules, see VanTol (2000).

Units that had not responded or had answered incorrectly were subject to mail, telephone and fax follow-up to ensure the data was obtained or corrected if needed. Also, the Internet was used to identify if

obtenir des estimations à ce niveau. Dans le cadre de l'EDI, l'établissement est choisi puisqu'il s'agit du niveau où l'information concernant les investissements est généralement disponible. Par contre, dans le cadre de l'ECET, la gestion des services électroniques et le commerce électronique sont habituellement maintenus au niveau de l'entreprise.

Comme le plan de sondage de l'EDI est au niveau établissement, nous avons utilisé la méthode du partage des poids (voir Lavallée, 1995) afin de produire des estimations au niveau entreprise pour l'ECET. La méthode du partage des poids est élaborée au point 3.5.11 – Estimation.

Les données de l'ECET ont été collectées au niveau de l'entreprise. Pour ce faire, nous avons dû dériver les entreprises reliées aux établissements échantillonnés par l'EDI. Généralement, pour chaque établissement du RE, on retrouve un numéro d'entreprise associé. Dans les autres cas, des établissements ont été groupés de sorte qu'il soit facile pour le répondant de fournir l'information voulue et que le groupement s'apparente à la définition d'une entreprise.

De plus, pour les besoins de la collecte et de l'imputation, nous avons calculé la variable auxiliaire au niveau de l'entreprise, soit en utilisant directement l'information sur le RE, soit en additionnant l'information disponible au niveau de l'établissement. La variable auxiliaire est le revenu pour les unités du secteur privé et le nombre d'employés pour les unités du secteur public. Or, pour certaines unités telles les unités des bureaux-chefs et des unités du secteur public, on a dû recourir à d'autres sources pour obtenir l'information auxiliaire. Nous avons utilisé des sources internes de Statistique Canada, notamment des données de l'Enquête sur l'emploi, la rémunération et les heures de travail et des données de la Division des institutions publiques. Finalement, l'imputation par la moyenne à l'intérieur de groupes homogènes a permis d'obtenir une variable auxiliaire pour les cas non-résolus.

5. Collecte et vérification des données

L'échantillon tiré au départ pour le compte de l'EDI avait déjà été utilisé et par le fait même on savait déjà que certains établissements étaient fermés, amalgamés ou tout simplement duplicatas. Pour le reste des unités, un questionnaire a été envoyé par la poste au niveau de l'entreprise invitant le répondant à le retourner dûment rempli.

À la saisie des données, des règles de vérifications sont appliquées à chaque questionnaire, telles des règles de cohérence. Par exemple, certaines règles vérifient que si certains champs sont codés, tous ceux qui y sont reliés sont également codés. Pour plus de détails sur les règles de vérification, consulter VanTol (2000).

Les unités n'ayant pas répondu ou ayant répondu incorrectement ont fait l'objet d'un suivi postal, téléphonique ou par fax pour s'assurer d'obtenir leurs réponses ou encore de les corriger au besoin. De plus, Internet a été utilisé afin d'identifier si certaines entreprises possédaient un site Web. Enfin, nous avons priorisé les suivis en tenant compte de la taille de l'entreprise, de l'importance des variables manquantes et du type d'incohérences sur le questionnaire.

certain enterprises had a web site. Finally, we prioritized the follow-ups by taking into account the size of the enterprise, the importance of the missing variables and the kind of inconsistencies on the questionnaire.

The response rate of the survey is not straightforward to calculate. Here are some statistics that show the criteria that were chosen to calculate it:
Original sample: 21,865 enterprises
Sample with questionnaire answered: 14,925 enterprises
Global response rate: 77%

Note that a number of received questionnaires are not usable or are not complete. Hence, we can conclude that the response rate by question might be different than the global one.

6. Non-response Study

A small study was performed to see if there was a bias due to non-response. With the subject of the survey, we wondered if enterprises who did not tend to respond were the ones not using electronic services. If this was the case, only using the respondents to estimate for the population would bring a bias by overestimating the use of electronic services.

We drew a systematic sample of about 250 non-respondents from all industries and sizes. Selected units were contacted by phone and were asked to answer four key questions, which are the use of Internet, the use of a web site, selling through the Internet and purchasing through the Internet.

The results were compared with the respondents' answers. A Khi2 test was performed on final data and there was no statistical evidence that the distributions were not the same. So we were able to adjust the respondents' weight to take the non-respondents into account without the fear of creating a bias (see section 3.5.11 for information on sampling weight adjustment). For more details on the non-response study, see Duval (2000).

7. Outlier Detection

Before imputation was done, we proceeded with outlier detection. Outlier detection was only done on the value of sales over the Internet and the auxiliary variable. Since there were few units reporting Internet sales, outlier detection for that variable was made within only 2 groups: public sector and private sector. For the auxiliary variable, detection was made within each industry. However, if there were less than 10 units available in an industry, detection was made within public sector and private sector groups. A method using the distance between observations was used (Nobrega, 1997). This method states that an observation Y_i is considered an outlier if, once enterprises within a group were ordered by the variable of interest Y, we have:

Le taux de réponse de l'enquête n'est pas simple à calculer. Voici quelques statistiques qui montre les critères qui ont été retenus pour le calculer :
Échantillon original : 21,865 entreprises
Échantillon avec questionnaire répondu : 14,925 entreprises
Taux de réponse global : 77%

Il faut noter que certains questionnaires reçus ne sont pas utilisables ou encore ne sont pas complets. On peut donc en conclure que le taux de réponse par question peut être différent au taux de réponse global.

6. Étude de non-réponse

Une étude a été réalisée afin d'évaluer le biais dû à la non-réponse. En effet, étant donné la nature de l'enquête, on s'est demandé si les entreprises qui n'avaient pas tendance à répondre étaient celles qui n'utilisent pas les services électroniques. Si c'était le cas, en n'utilisant que les répondants pour inférer à la population, on obtiendrait un biais en surestimant l'utilisation des services électroniques.

Dans le cadre de l'étude, un échantillonnage systématique d'environ 250 entreprises de différentes tailles et de différentes industries a été sélectionné parmi les non-répondants. Les unités sélectionnées ont été contactées par téléphone et invitées à répondre à quatre questions clés soient : l'utilisation d'Internet, l'utilisation d'un site Web, la vente par le biais d'Internet et l'achat par le biais d'Internet.

Les résultats obtenus ont été comparés à ceux de l'ensemble des répondants. Un test du Khi2 a été utilisé sur les données finales et aucune différence statistique importante entre les distributions des répondants et des non-répondants n'a été signalée. On a donc pu, à l'estimation, repondérer les répondants pour tenir compte des non-répondants en utilisant le plan de sondage initial sans craindre de créer un biais (voir la section 3.5.11 pour de l'information sur la repondération). Pour plus de détails sur l'étude de non-réponse voir Duval (2000).

7. Détection de données aberrantes

Avant de procéder à l'imputation, on a commencé par effectuer la détection de données aberrantes. La détection s'est fait sur deux variables : la valeur des ventes sur Internet et la variable auxiliaire. Pour la valeur des ventes, comme il y avait peu d'unités faisant des ventes par Internet, la détection s'est fait à l'intérieur de 2 groupes seulement : le secteur public et le secteur privé. Pour la variable auxiliaire, la détection s'est fait dans chacune des industries. Cependant, si moins de 10 entreprises étaient disponibles pour une ou des industries, la détection a été effectuée à l'intérieur des groupes du secteur privé et du secteur public. Une méthode basée sur les écarts entre les observations a été utilisée (Nobrega, 1997). Cette méthode considère qu'une donnée Y_i est aberrante si, une fois les entreprises d'un groupe trié en ordre croissant selon la variable d'intérêt Y, on a :

$$Y_i - Y_{i-1} > M_D + 3*SVD_D$$
$$Y_i > M_D$$

$$Y_i - Y_{i-1} > M_D + 3*SVD_D$$
$$Y_i > M_D$$

Where Y_{i-1} is the value of Y for observation i-1,
SVD_D is group D's standard deviation,
M_D group D's median.

As soon as a value was detected as an outlier, greater values were also outliers.

Then, outliers were analyzed and corrected if necessary. Outlier enterprises were not used during imputation, they were excluded from the donor pool. However, if outlier enterprises needed imputation, we imputed them, but we kept outlier values.

Moreover, if some of the establishments found to be outliers formed part of the take-some stratum and did not correctly represent other units in the stratum, they were moved up to the take-all stratum. That way, they would only represent themselves. The selection probability for residual units was then recomputed.

8. Imputation

We first tried to correct or complete questionnaires through follow-up of respondents, but when the survey collection was closed and some records were incomplete, we proceeded with imputation in order to correctly fill them and to ensure consistency. Amongst all records to be imputed, there were some that were incomplete (but partially filled out), some that had invalid response patterns and finally some that did not satisfy edit rules.

Many imputation methods were used: imputation using administrative data, historical imputation and donor imputation. Every record was identified and completed in order to change the respondents' answers by the least amount possible.

In the case of total non-response, no imputation was performed and we simply reweighted responding units under the assumption that non-response was randomly distributed (see section 3.5.8: Non-response Study). In some sense, we acted as if the unit had never been selected and the weights were adjusted in order to account for it. The target precision of the estimate is then decreased.

Deterministic imputation was used when answers from questions related to the question needing imputation lead to only one possible answer that would maintain the questionnaire's consistency. This unique value was then imputed.

Imputation using administrative data was used to impute the question referring to the number of employees by using the number of employees available in the BR or other sources.

où Y_{i-1} est la valeur de Y de l'observation i-1,
SVD_D est l'écart-type des unités du groupe D,
M_D est la médiane des unités du groupe D.

Dès qu'une valeur est détectée aberrante, les valeurs supérieures sont aussi considérées aberrantes.

Ces données ont ensuite été vérifiées et corrigées au besoin. Les enregistrements trouvés aberrants n'ont pas été utilisés dans les calculs lors de l'imputation. Par exemple, pour l'imputation par donneur, on les a exclus du bassin de donneurs. Par contre, s'il s'agissait des entreprises à être imputées, on a procédé à l'imputation de celles-ci, tout en conservant leurs valeurs aberrantes validées.

Aussi, si certaines des unités trouvées aberrantes étaient considérées mal classifiées lors de l'échantillonnage et ne représentaient pas correctement les autres unités de la strate, alors elles étaient promues dans une strate à tirage complet. Ainsi, elles ne représentaient qu'elles-mêmes. La probabilité de sélection des unités résiduelles a été recalculée.

8. Imputation

On a tout d'abord tenté de corriger les questionnaires ou encore de les compléter, en faisant un suivi auprès du répondant, mais lorsque l'enquête a été fermée et qu'il restait certains enregistrements toujours incomplets, il a fallu procéder à l'imputation pour permettre de les remplir correctement et de façon cohérente. Parmi les enregistrements à imputer, on retrouvait tous ceux qui étaient incomplets (mais partiellement remplis), ceux dont les patrons de réponse ne suivaient pas un patron valide et finalement ceux qui ne satisfaisaient pas les règles de vérification.

Plusieurs types d'imputation ont été utilisés, soit l'imputation par source administrative, l'imputation historique et l'imputation par donneur. Chacun des enregistrements a été identifié et complété de façon à changer le moins possible les données du répondant.

Dans le cas d'une non-réponse totale, aucune imputation n'a été faite et on s'est contenté tout simplement de repondérer les unités répondantes sous l'hypothèse que la non-réponse était aléatoire (voir section 3.5.8: Étude de non-réponse). La repondération consiste en gros à ajuster les poids de sondage initiaux en supposant que les unités non-répondantes étaient des unités non-sélectionnées au départ. La précision visée à l'estimation est alors amoindrie.

L'imputation déterministe a été effectuée lorsque les réponses aux questions reliées à la question à imputer ne laissaient qu'un seul choix de réponse permettant le maintien de la cohérence du questionnaire. Cette seule valeur possible a été alors utilisée.

L'imputation par source administrative a été effectuée pour la question portant sur le nombre d'employés en utilisant le nombre d'employés disponible sur le registre des entreprises et autres sources administratives.

L'imputation historique a été utilisée pour imputer certaines questions lorsque l'entreprise avait répondu positivement à la question (en d'autres mots, qu'elle utilisait la technologie concernée) l'an dernier. On a supposé qu'une entreprise ayant

Historical imputation was used to impute some questions when the enterprise positively responded to the question (in other words, it used the concerned technology) last year. We assumed that an enterprise positively responding last year should still respond positively this year. However, if the enterprise negatively responded (i.e. it did not use that technology) last year, we would not use that information to impute, since it is possible that the situation has changed since last year.

Finally, in the remaining cases, **donor imputation** was used to replace missing or incoherent values with those of the nearest respondent according to characteristics such as size, industrial classification and key variables from the questionnaire. We also checked to be sure that the imputed values did not affect the questionnaire's consistency. If it did, we used another donor.

For donor imputation purposes, we divided the questionnaire into 2 independent groups of variables. The first group contained variables concerning use of electronic services and Internet sales. The second group focused on technological and organizational improvements. These groups were imputed in separate ways. A record completed for one section (without having a totally filled questionnaire) became a donor. One record can then be imputed from two different donors. The advantage of doing this was to maximize the number of potential donors per section.

Imputation was conducted within homogeneous groups. The initial imputation group corresponded to NAICS-4 level and size grouping. In the case of the public sector, the size grouping was based on the number of employees:
Group Size1) less than 100 employees
Size2) 100 employees and more
In the private sector, the size grouping was based on the revenue:
Group Size1) less than $10,000,000
Size2) $10,000,000 and more

Note that outlier enterprises were excluded from the donor pool. If there were not at least 10 potential donors and 25% of donors in a group, or if imputation from all available donors would result in questionnaire inconsistencies, we moved to a more aggregated imputation group. We aggregated within the following order:
NAICS-4 level;
NAICS-3 level and size grouping;
NAICS-3 level;
NAICS-2 level and size grouping;
NAICS-2 level.

When we could not find any donor for an enterprise, it was manually imputed. This situation rarely happened. When imputation was done, we adjusted the sales value over the Internet by the ratio of imputed and donor's revenue.

Finally, when imputation was over, we reapplied the

répondu positivement l'an dernier devrait encore répondre par l'affirmative cette année. Cependant, si l'entreprise avait répondu négativement à la question (qu'elle n'utilisait pas la technologie) l'an dernier, cette information n'a pas été utilisée pour imputer car il est fort possible que la situation ait changé depuis l'année dernière.

Finalement, dans les autres cas, **l'imputation par donneur** a été effectuée en remplaçant les valeurs manquantes ou incohérentes par celles du plus proche répondant selon certaines caractéristiques telles la taille, la classification industrielle et les variables-clé du questionnaire. De plus, on s'est assuré que le donneur permettait de respecter la cohérence entre les champs imputés et les champs rapportés du receveur. Sinon, un autre donneur a été utilisé.

Pour fins d'imputation par donneur, on a divisé le questionnaire en 2 groupes de variables, les 2 groupes formés étant indépendants entre eux. Le premier groupe portait sur les variables concernant l'utilisation des services électroniques et la vente par Internet, alors que le deuxième groupe portait sur les améliorations technologiques et organisationnelles. Chaque groupe a été imputé séparément. Un enregistrement complet pour une section (sans être complet pour l'ensemble des questions) devenait alors donneur. On peut donc avoir deux donneurs différents pour un même questionnaire. L'avantage était de maximiser le nombre de donneurs potentiels par section.

L'imputation a été exécutée à l'intérieur de groupes homogènes. Le groupement initial correspondait au SCIAN de niveau 4, divisé en 2 tailles :
Dans le cas du secteur public, celle-ci était basée sur le nombre d'employés :
Groupe Taille1) moins de 100 employés;
Taille2) 100 employés et plus.
Alors que dans le cas du secteur privé, celle-ci était basée sur le revenu :
Groupe Taille1) moins de 10 000 000$;
Taille2) 10 000 000$ et plus.

Notons que les questionnaires avec données aberrantes étaient exclus du bassin de donneurs. De plus, si on n'avait pas au moins 10 donneurs potentiels et 25% de donneurs par groupe ou encore, si aucun donneur disponible ne permettait l'imputation tout en respectant les règles de validation du questionnaire receveur, on passait à un groupe d'imputation plus agrégé. On a agrégé dans l'ordre suivant:
SCIAN de niveau 4;
SCIAN de niveau 3 et les groupes de taille;
SCIAN de niveau 3;
SCIAN de niveau 2 et les groupes de taille;
SCIAN de niveau 2.

Dans les cas où on ne pouvait pas trouver de donneurs, ces unités ont été imputées manuellement. Cette situation était très rare. Une fois l'imputation effectuée, on a ajusté la variable des ventes par Internet par le ratio des revenus du receveur et du donneur.

Enfin, une fois l'imputation terminée, les règles de vérification initiales ont été réappliquées afin de s'assurer de la cohérence de tous les questionnaires utilisés à l'estimation. Des drapeaux d'imputation ont été créés afin de garder l'information des variables imputées.

initial edit rules to assure the consistency of all the questionnaires going into the estimation process. Imputation flags were created to keep information about imputed fields.

9. Estimation

Statistics Canada's Generalized Estimation System (GES) was used (see Esteavo, 1991). The estimation was done in two phases: the first phase sample was the initial sample and the second phase sample was the respondents' sample. The groups used in both phases were the original strata (with changes, as previously explained) since there was no bias due to non-response.

We have first phase weighting: $W_{h1}=N_h/n_h$
and second phase weighting: $W_{h2}=n_h/n_h'$
where N_h is the initial population per stratum, n_h, the initial sample per stratum and n_h', the respondents' sample.

We can then calculate the unit's final weight by $Wh=W_{h1}$ X $W_{h2}=N_h/n_h'$.

As stated previously, we want to produce estimates at the enterprise level since the relevance and the availability of the use of electronic services and electronic commerce is obtained at that level. Unfortunately, the CES sampling was performed at the establishment level. To solve this problem, we used the weight share method (see Lavallée, 1995). This method is used when the targeted population has to be sampled using a survey frame corresponding to a different population, but linked in a certain way to the first one. This is the case in our situation because every establishment is associated with an enterprise. The estimator then obtained is unbiased.

The following figure shows the establishment population (represented by white circles) U^A, the enterprise population (represented by grey rectangles) U^B and the links between an establishment and the associated enterprise. Obviously, each establishment is linked to one and only one enterprise.

Figure 3 : Links between establishment and enterprise populations

9. Estimation

Le système généralisé d'estimation (SGE) de Statistique Canada a été utilisé (voir Esteavo, 1991). L'estimation s'est fait en deux phases : l'échantillon de première phase étant l'échantillon initial et l'échantillon de deuxième phase, l'échantillon de répondants. La même stratification a été utilisée en première et deuxième phases étant donné l'absence de biais dû à la non-réponse.

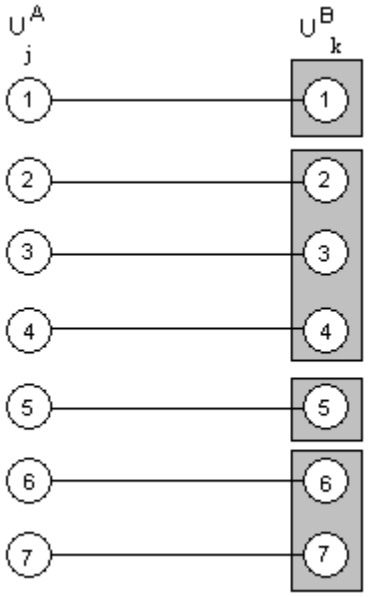
On a la pondération de première phase : $W_{h1}=N_h/n_h$
et la pondération de deuxième phase : $W_{h2}=n_h/n_h'$
où N_h est la population de départ par strate, n_h, l'échantillon de départ par strate et n_h' l'échantillon de répondants.

On peut donc réécrire le poids final de chaque unité comme étant $W_h=W_{h1}$ X $W_{h2} = N_h/n_h'$.

Comme déjà mentionné, on veut obtenir des estimés au niveau entreprise puisque c'est à ce niveau que l'information concernant l'utilisation des services électroniques et le commerce électronique est disponible et pertinente. Or, l'échantillonnage de l'EDI s'est fait au niveau de l'établissement. Pour contrer ce problème, on a eu recours à la méthode du partage des poids (voir Lavallée,1995). Cette méthode est utilisée lorsque la population visée doit être échantillonnée en vertu d'une base de sondage correspondant à une population différente, mais lié d'une certaine façon à la première. Ce qui est le cas dans notre situation puisque chaque établissement est associé à une entreprise. L'estimateur ainsi obtenu est non-biaisé.

La figure suivante montre la population des établissements (représentés par des cercles blancs) U^A, la population des entreprises (représentées par des rectangles gris) U^B et les liens reliant un établissement à l'entreprise auquel il appartient. Évidemment, chaque établissement n'est relié qu'à une et une seule entreprise.

Schéma 3 : Liens entre la population des établissements et celle des entreprises



The weight share method assigns to each establishment k (of enterprise i) of the CES population an initial weight w'_{ik} calculated as follows :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$$

where : $l_{j,ik}$ = 1 if a link exists between unit j and unit k, 0 otherwise,
t_j = 1 if unit j is sampled, 0 otherwise,
π_j^A is unit j's selection probability,
M^A is the number of units in U^A.

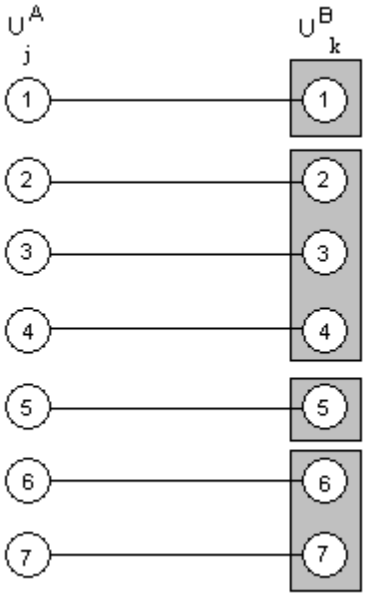
The enterprise i weight (w_i) is obtained by the formula

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}}$$

where : L_{ik} is the number of links between unit k and population U^A (this number always equals 1 in our case),
M_i^B is the number of establishments in population U^B.

The value $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ represents the number of establishments in enterprise i.

We assign $w_{ik} = w_i$ as the establishment's final weight i.e. every establishment within an enterprise has the same average weight.



La méthode du partage des poids attribue à chaque établissement k (de l'entreprise i) de la population de l'EDI un poids initial w'_{ik} calculé comme suit :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A}$$

où : $l_{j,ik}$ = 1 s'il existe un lien entre l'unité j et l'unité k, 0 sinon,
t_j = 1 si l'unité j est échantillonnée, 0 sinon,
π_j^A est la probabilité de sélection de l'unité j,
M^A est le nombre d'unités dans U^A.

Le poids de l'entreprise i (w_i) s'obtient par la formule

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}}$$

où : L_{ik} est le nombre de liens entre l'unité k et la population U^A (ce nombre est toujours égal à 1 dans notre cas),
M_i^B est le nombre d'établissements dans la population U^B.

La valeur $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ représente alors le nombre d'établissements dans l'entreprise i.

On attribue $w_{ik} = w_i$ comme étant le poids final de l'établissement i.e. tous les poids des établissements échantillonnés d'une même entreprise ont le même poids moyen.

Des estimations ont été produites pour toutes les variables et

Estimates were produced for all variables and all desired publishing domains, which generally are different NAICS levels (NAICS-2 to NAICS-4 levels depending of the industrial classification of interest). Here are the estimation formulas used depending on the type of variable that was being estimated.

1) In the case of percentage variables (*P*), a ratio has been used to derive an estimate.

$$\hat{P}_d = \frac{\sum_s w_i z_i p_i(d)}{\sum_s w_i z_i} \text{ where } p_i(d) = \begin{cases} p_i \text{ if } i \varepsilon d \\ 0 \text{ otherwise} \end{cases}$$

The variable z is the auxiliary variable which may be revenue or the number of employees depending on the variable being estimated. This variable, if used, allows us to produce economically weighted estimates. The variable w is the final weight of unit i after reweighting and calibration. For more details on the calculation of the weight w, see Arcaro (1998).

2) In the case of categorical variables (*C*), again a ratio has been used.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \text{ where}$$

$$c_i(d) = \begin{cases} 1 \text{ if } i \varepsilon d \text{ and the category was chosen} \\ 0 \text{ otherwise} \end{cases}$$

The variable z is the auxiliary variable which may be revenue or the number of employees depending on the variable being estimated.

3) In the case of numerical variables (*Y*), the usual estimator of the total is used.

$$\hat{Y}_d = \sum_s w_i y_i(d) \text{ where } y_i(d) = \begin{cases} y_i \text{ if } i \varepsilon d \\ 0 \text{ otherwise} \end{cases}$$

For formulae for variance estimation for each type of variable (*P*, *C* and *Y*), please refer to Arcaro (1998).

Calculation of CV

The coefficient of variation (CV) is computed using the ratio:

$$CV(\hat{Y}(d)) = \frac{\sqrt{\hat{V}(\hat{Y}(d))}}{\hat{Y}(d)}$$

where the numerator represents the estimate's standard deviation. Variable Y may represent any of the types of variables already discussed. However, in cases of percentage or categorical variables, we modified the

tous les domaines de publication désirés qui sont en général différents niveaux de SCIAN (niveau 2 au niveau 4 dépendamment des classifications industrielles d'intérêt). Voici les formules d'estimation utilisées pour chaque domaine selon la catégorie de variable visée.

1) Dans le cas des variables de pourcentage (*P*), un quotient a été utilisé pour produire les estimations.

$$\hat{P}_d = \frac{\sum_s w_i z_i p_i(d)}{\sum_s w_i z_i} \text{ où } p_i(d) = \begin{cases} p_i \text{ si } i \varepsilon d \\ 0 \text{ si non} \end{cases}$$

La variable w représente le poids final de l'unité i après repondération et partage des poids. La variable z, est une variable auxiliaire qui peut être le revenu ou le nombre d'employés selon le pourcentage estimé. Des estimés sont produits avec et sans cette variable auxiliaire. Cette variable permet de dériver des estimés qu'on appelle économiquement pondérés en donnant plus de poids aux unités de grandes tailles.

2) Dans le cas des variables catégoriques (*C*), à nouveau un quotient a été utilisé.

$$\hat{C}_d = \frac{\sum_s w_i z_i c_i(d)}{\sum_s w_i z_i} \text{ où}$$

$$c_i(d) = \begin{cases} 1 \text{ si } i \varepsilon d \text{ et la catégorie a été choisie} \\ 0 \text{ si non} \end{cases}$$

La variable z, est une variable auxiliaire qui peut à nouveau être le revenu ou le nombre d'employés selon la variable estimée, ou encore ne pas être utilisé.

3) Dans le cas des variables numériques (*Y*), l'estimateur habituel du total est utilisé

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

Pour ce qui est des formules d'estimation de variance pour chacune des catégories de variable (*P*,*C* et*Y*), il faut se référer à Arcaro (1998).

Calcul du CV

Le coefficient de variation (CV) est calculé à l'aide du quotient:

$$CV(\hat{Y}(d)) = \frac{\sqrt{\hat{V}(\hat{Y}(d))}}{\hat{Y}(d)}$$

où le numérateur représente l'écart-type échantillonnale de l'estimation. La variable Y peut représenter chacun des types de variables discutés préalablement. Par contre, dans le cas de pourcentages et de variables catégoriques, on a modifié le calcul du CV en utilisant Y(d)=0.5. On évite ainsi d'obtenir de très petits

CV calculation by using Y(d)=0.5. This way, we avoid getting very small or very large CVs due to Y(d) being close to 1 or close to 0.

This coefficient tries to give a relative measure of the error made when using a sample instead of using a census to derive an estimate about the whole population.

Confidentiality

Some confidentiality rules are used to suppress any information that might lead to disclosure of the data supplied by a respondent. These rules allow Statistics Canada to comply with its mandate of non-disclosure of information supplied by respondents. The rules themselves are confidential and are not available for consultation.

10. Sampling error and non-sampling error

The difference between an estimate based on sample data and the value obtained by surveying the entire population is called the sampling error. This difference varies with sample size, variability of the variable of interest, sampling design, and estimation method. In general, the larger a sample, the smaller its sampling error. If the population is very heterogeneous, a larger sample size is required to produce a reliable estimate. The sampling error is measured by a quantity known as the standard deviation. The latter indicates the expected variability of the estimate that will be produced if we sample repeatedly. The actual value of the standard deviation is unknown, but it can be estimated from the sample.

Another measure of precision is the coefficient of variation (CV). The CV is simply the standard deviation expressed as a percentage of the estimate. Hence it is a relative measure of precision and can be used for comparisons across industries or provinces. The smaller the CV, the more reliable the estimate.

Another kind of error is non-sampling errors such as frame problems, response errors, data capture errors, etc. Although every effort is made to keep such errors to a minimum, they always exist. They are not taken into account in computing the CV, nor are they measured by the CV. Measures such as response rate, coverage rate and imputation rate can be used as indicators of the possible extent of non-sampling errors.

When the estimates are published, a scale distinguishes between the various qualities of accuracy. It combines the effect of sampling (using the CV) and the imputation rate (each imputed value adds to the uncertainty of the results). The scale is presented in Table 6.

ou grands CV reliés au fait que Y(d) soit très près de 1 ou très près de 0.

Ce coefficient tente de donner une mesure relative de l'erreur commise lorsqu'on a recours à un échantillon au lieu de produire une statistique à l'aide de l'ensemble de la population.

Confidentialité

Certaines règles de confidentialité sont utilisées pour supprimer toute information qui pourrait mener à la divulgation des données fournies par un répondant. Ces règles permettent à Statistique Canada de respecter son mandat de non-divulgation d'information fournie par les répondants. Les règles elles-mêmes sont confidentielles et ne sont pas disponibles pour consultation.

10. Erreur d'échantillonnage et non-due à l'échantillonnage

La différence entre l'estimation produite à partir de données échantillonnées et de données recensées est appelée erreur d'échantillonnage. Cette différence varie plus ou moins selon la taille de l'échantillon, la variabilité de la variable d'intérêt, le plan de sondage et la méthode d'estimation. En général, un échantillon plus grand produit une erreur d'échantillonnage plus petite. Si la population est très hétérogène, une taille d'échantillon plus grande est requise pour produire une estimation fiable.

L'erreur d'échantillonnage est mesurée par une quantité appelée écart-type. Cette quantité mesure la variabilité anticipée de l'estimation produite si on fait un échantillonnage répété. La vraie valeur de l'écart-type est inconnue mais peut être estimée à partir de l'échantillon.

Une deuxième mesure de précision est le coefficient de variation (CV). Ce coefficient est simplement l'écart-type exprimé en pourcentage de la valeur de l'estimation. Il donne donc une mesure de précision relative et comparable entre différentes industries ou provinces. Notons qu'un plus petit CV indique une plus grande fiabilité de l'estimation.

En plus de l'erreur d'échantillonnage, il existe des erreurs non-dues à l'échantillonnage telles des problèmes de base de sondage, des erreurs de réponses, des erreurs lors de l'encodage des réponses, etc., sur lesquelles on tente de conserver un contrôle des plus stricts. Néanmoins, celles-ci existent toujours et ne sont pas comptabilisées lorsque l'on calcule le coefficient de variation et ne sont pas incluses dans cette mesure de précision. Certaines mesures telles que les taux de réponses, de couverture et d'imputation peuvent être utilisées comme indicateurs du niveau potentiel des erreurs non-liées à l'échantillonnage.

Lors de la publication des estimations, une échelle permet de distinguer entre les différentes qualités de précision. Celle-ci combine l'effet dû à l'échantillonnage (à l'aide du CV) et le taux d'imputation (chaque imputation ajoute à l'incertitude des résultats). L'échelle utilisée est reproduite au tableau 6.

Table 6
Quality indicator interpretation

CV	Imputation rate			
	0.00 - 0.10	0.10 - 0.33	0.33 - 0.60	0.60 - +++
0.00 - 0.05	A	B	C	F
0.05 - 0.10	B	C	D	F
0.10 - 0.15	C	D	E	F
0.15 - 0.25	D	E	F	F
0.25 - 0.50	E	F	F	F
0.50 - +++	F	F	F	F

A: Excellent B: Very good C: Good
D: Acceptable E: Use with caution F: Unpublishable

Unpublishable data should be hidden while data that is considered "use with caution" should be identified as such.

11. References

Arcaro (1998). GES Estimation Specifications for Two-Phase Sampling with Auxiliary Information, Internal Statistics Canada document, 1998.

Cuthill, I. (1996). The Statistics Canada Business Register. Internal Statistics Canada Document, 1996.

Duval, Landry (2000). Étude de non-réponse pour l'enquête sur le commerce électronique 2000. , Internal Statistics Canada document, May 2001.

Estevao, V. (1991). Generalized Estimation System, Methodology Review. Internal Statistics Canada document, September 1991.

Hidiroglou, M.A. (1986). The Construction of a Self-Representing Stratum of Large Units in Survey Design. The American Statistician, 40, 27-31

Latouche, M. (1988). Détermination, allocation et sélection de l'échantillon. Cahier de travail BSMD-88-021 de Statistique Canada, May 1988

Nobrega, Karla (1998). Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method. SSC, 1998 Proceedings of the Survey Methods Section, June 1998.

Pandher G.H. (1995). Population asymétrique: construction optimale de groupes "à tirage complet" et "échantillons", avec application au remaniement de l'enquête sur les finances des administrations locales. Cahier de travail SSMD-95-001 de Statistique Canada, March 1995.

Pierre Lavallée (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, Survey Methodology, volume 21, june 1995, 25-32.

Tableau 6
Interprétation de la côte de qualité

CV	Taux d'imputation			
	0.00 - 0.10	0.10 - 0.33	0.33 - 0.60	0.60 - +++
0.00 - 0.05	A	B	C	F
0.05 - 0.10	B	C	D	F
0.10 - 0.15	C	D	E	F
0.15 - 0.25	D	E	F	F
0.25 - 0.50	E	F	F	F
0.50 - +++	F	F	F	F

A: Excellent B: Très bon C: Bon
D: Acceptable E: Utiliser avec réserve F: Non-publiables

Les données non-publiables devraient être cachées alors que les données « utilisable avec réserve » devraient être identifiées comme telles.

11. Références

Arcaro (1998). GES Estimation Specifications for Two-Phase Sampling with Auxiliary Information, Document interne de Statistique Canada, 1998.

Cuthill, I. (1996). The Statistics Canada Business Register. Document interne de Statistique Canada, 1996.

Duval, Landry (2000). Étude de non-réponse pour l'enquête sur le commerce électronique 2000. Document interne de Statistique Canada, mai 2001.

Estevao, V. (1991). Generalized Estimation System, Methodology Review. Document interne de Statistique Canada, septembre 1991.

Hidiroglou, M.A. (1986). The Construction of a Self-representing Stratum of Large Units in Survey Design. The American Statistician, 40, 27-31.

Latouche, M. (1988). Détermination, allocation et sélection de l'échantillon. Cahier de travail BSMD-88-021 de Statistique Canada, mai 1988.

Nobrega, Karla (1998). Outlier Detection in Asymmetric Samples: A Comparison of an Inter-quartile Range Method and a Variation of a Sigma Gap Method. SSC, 1998 Proceedings of the Survey Methods Section, June 1998.

Pandher G.H. (1995). Population asymétrique: construction optimale de groupes "à tirage complet" et "échantillons", avec application au remaniement de l'enquête sur les finances des administrations locales. Cahier de travail SSMD-95-001 de Statistique Canada, mars 1995.

Pierre Lavallée (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. Techniques d'enquêtes, volume 21, juin 1995, 27-35.

VanTol, Bryan (2000). Edits2000 Internal Statistics Canada document, December 1995.

VanTol, Bryan (2000). Edits2000. Document interne de Statistique Canada, décembre 1995.