Techniques d'enquête, juin 1989
Vol. 15, n° 1, pp. 151–165
Statistique Canada

Utilisation de données administratives pour l'établissement des profils initiaux et ultérieurs des entités économiques

COLLEEN CLARK et ROBERT LUSSIER[1]

RÉSUMÉ

Statistique Canada s'emploie actuellement à remanier son registre central des entités économiques. Dans le nouveau registre, chaque entité économique est considérée comme un réseau d'entités juridiques et exploitantes dont les caractéristiques permettent de déterminer des entités statistiques. On obtient l'image de ce réseau, c'est-à-dire le profil par le processus dit d'«établissement des profils», qui suppose des contacts avec l'entité économique. En 1986, on s'est servi d'une liste de toutes les entités avec lesquelles il fallait entrer en contact afin d'obtenir les profils permettant de constituer le nouveau registre. Pour dresser cette liste, on a eu recours à des données administratives. À l'avenir, les données administratives serviront de source de renseignements sur les changements qui auront pu se produire dans les entités économiques. Elles pourront donc être utilisées comme source de mise-à-jour directe ou comme signal qu'on doit reviser la structure d'une entité. L'article porte d'abord sur les objectifs du processus d'établissement des profils. On présente ensuite les procédures de construction de la base de sondage servant au processus d'établissement des profils initiaux à l'aide de plusieurs sources de données administratives. Ces procédures comprennent l'application de concepts, la détection des cas de chevauchement entre les sources et l'évaluation de la qualité des données. On examine ensuite le rôle des données administratives comme source de renseignements sur les changements qui peuvent s'être produits dans les entités économiques et comme source de données sur lesquelles on peut s'appuyer pour demander de vérifier les profils. Suit une analyse des résultats d'une étude de simulation visant à évaluer ce rôle. L'exposé s'achève par une série de questions sur la méthodologie relative à l'utilisation de données administratives en vue de l'établissement de profils de mise à jour.

MOTS CLÉS: Données administratives; registre central; profil.

1. INTRODUCTION

Statistique Canada s'emploie actuellement à réorganiser son programme d'enquêtes économiques. Dans le nouveau programme, on utilisera davantage les données administratives. Ces dernières feront partie intégrante d'une Base de Données du Registre Central (BDRC) d'où les enquêtes économiques tireront leur échantillons.

Les données administratives serviront aussi à mettre à jour la BDRC. Cet aspect et d'autres aspects de la stratégie de réorganisation sont présentés dans Colledge et Lussier (1985). Certains résultats de la mise en pratique de la stratégie sont contenus dans Colledge (1987).

Une des premières étapes a été de définir les unités de la BDRC. L'unité fondamentale est l'entité commerciale. Statistique Canada (voir Statistique Canada, 1987) définit une entité commerciale comme: «un agent économique ayant la responsabilité et le pouvoir d'affecter des ressources à la production de biens et/ou de service et, de ce fait, de gérer et de contrôler la façon de recevoir et d'employer les recettes, d'accumuler des avoirs, d'emprunter et de prêter des capitaux et de tenir à jour des états financiers complets de ses activités» (traduction).

La Base de Données du Registre Central actuellement élaborée par Statistique Canada est une entreprise de représentation de la structure de l'économie canadienne. Elle tient compte

[1] Colleen Clark, Division des méthodes d'enquêtes sociales, Statistique Canada, 4-C1, Immeuble Jean Talon, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6; Robert Lussier, Division des méthodes d'enquêtes-entreprises, 11-M, Immeuble R.H. Coats, Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6.

du fait que cette économie est dominée par un petit nombre de grandes entités commerciales qui représentent le gros de l'activité dans cette économie. La BDRC est divisée en deux parties pour traduire cette dichotomie.

Une des composantes, la partie intégrée (PI), couvre le petit nombre d'entités commerciales importantes par leur taille ou selon d'autres critères, tandis que l'autre composante, la partie non intégrée (PNI), couvre les autres entités commerciales, c'est-à-dire le grand nombre de petites entités. Les entités de la première composante sont plus complexes. Aussi faut-il un certain effort pour déterminer, dans une entité commerciale complexe, quels éléments peuvent être intéressants pour une enquête donnée.

La partie intégrée (PI) de la BDRC est un moyen de représentation de la structure complexe des entités commerciales à l'aide d'un modèle d'information. Le modèle se compose de cinq structures liées entre elles qui décrivent une entité commerciale. Ces structures permettent de définir avec précision les populations d'enquête. Les cinq structures sont les suivantes.

i. La *structure juridique*, qui décrit comment l'entité commerciale est organisée juridiquement. Elle représente les entités juridiques et les relations de propriété et de contrôle qu'elles ont entre elles. Entre autres exemples d'entités juridiques, notons les entreprises constituées en société en vertu de chartes fédérales ou provinciales.

ii. La *structure opérationnelle*, qui décrit comment l'entité commerciale fonctionne et comment elle organise son système comptable. Elle se compose des entités exploitantes. C'est elle qui organise et contrôle la production des biens et/ou des services. Il s'agit d'un moyen de structurer l'entité commerciale de la façon dont cette dernière se perçoit. Entre autres exemples d'entités exploitantes, notons les divisions, les centres de profit et les usines.

iii. La *structure statistique*, qui représente les entités statistiques classées selon un ordre hiérarchique. Les entités statistiques sont constituées à partir de la structure opérationnelle correspondante suivant les unités de la structure opérationnelle pour lesquelles un ensemble particulier de données est tenu à jour.

iv. La *structure déclarante*, qui représente les modalités de déclaration définies pour chacune des entités statistiques choisies, par enquête. Les données du système comptable de l'entité commerciale sont communiquées par les entités déclarantes.

v. La *structure administrative*, qui contient des données administratives comme les données fiscales recueillies auprès des entités juridiques et les données des comptes de retenue sur la paye recueillies auprès des entités exploitantes.

Les entités des structures statistiques et déclarantes sont produites par le Bureau à des fins de collecte, de vérification, d'estimation et de totalisation de données économiques. Les entités des trois autres structures sont définies par les organismes externes à Statistiques Canada.

Le processus complexe de délimitation des frontières de chacune des entités commerciales et de détermination de ses cinq structures et des relations existant entre ces structures est appelé «établissement de profils». Cette représentation de l'entité commerciale comme réseau est le «profil». Les données servant à construire un profil sont obtenues par contact avec l'entité commerciale ou avec une de ses composantes. Les éléments d'information sur la structure juridique et sur la structure opérationnelle des entités commerciales ainsi que certains éléments d'information sur leur structure administrative sont obtenues ou révisées et mis à jour au cours de l'interview. La structure statistique est ensuite produite ou mise à jour automatiquement à partir de la nouvelle structure opérationnelle. Finalement, des entités déclarantes implicites sont créées pour chaque entité statistique nouvellement choisie à l'aide des données de certaines zones tirées des structures juridique, opérationnelle ou administrative. Ces entités peuvent par la suite être mises à jour au moyen des renseignements obtenus à l'occasion du premier contact avec les répondants ou au moyen d'arrangements spéciaux négociés avec les répondants.

Le type de contact qui est pris en vue de l'établissement des profils dépend de la complexité des entités et de toute modalité de déclaration spéciale. Pour ce qui est des données relatives aux entités les plus complexes et les plus importantes, elles seront recueillies par le personnel

du bureau central ou d'un bureau régional par interview sur place. Les données relatives aux autres entités seront recueillies par interview téléphonique. Les entités seront contactées une fois tous les deux ans ou plus souvent, selon la rapidité avec laquelle leurs structures changent.

La méthode d'établissement cyclique des profils, par laquelle les entités commerciales sont contactées périodiquement, est une des méthodes qui seront utilisées pour mettre à jour la PI de la BDRC. On utilisera également des renseignements tirés d'enquêtes et des données de sources administratives.

La conception et la construction de la BDRC s'échelonne sur une période de trois ans qui devrait aboutir à la création d'une base de données à intégrer dans les programmes des enquêtes. Lors de la mise en application de cette base, la plupart des données relatives à la partie intégrée de la BDRC auront été obtenues au moyen du processus d'établissement des profils qui a commencé en avril 1986. Mais à ce moment-là, il n'y avait pas de liste unique d'entités commerciales dont on pouvait établir le profil.

Les données administratives ont joué un rôle important dans l'amorce du processus d'établissement des profils. Statistique Canada s'en est servi comme point de départ pour construire son registre des entités commerciales. Une liste des entités commerciales pouvant faire l'objet d'un profil initial a été dressée à partir de sources de données administratives. La partie 2 du présent document décrit comment cela s'est passé. La partie 2.1 présente les éléments nécessaires à la création de la base de sondage. Une description des sources de données utilisées pour construire la base suit dans la partie 2.2, tandis que la partie 2.3 montre comment on a déterminé l'unité de sondage et comment on a combiné les diverses sources de données pour construire la base de sondage.

La partie 3 décrit comment les données administratives seront utilisées pour détecter les changements qui se seraient produits dans une entité commerciale et pour lancer ensuite le processus de mise à jour des profils. Sont ensuites présentés les résultats d'une étude de simulation effectuée pour quantifier le degré d'utilisation proposé des sources de données administratives. Enfin, le document se termine par une analyse de certains points soulevés dans cette étude.

2. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DE PROFILS INITIAUX

2.1 Éléments nécessaires à la création de la base de sondage

La première étape de la construction de la base de sondage devant servir à l'établissement des profils initiaux a été de définir l'unité de sondage. L'unité idéale aurait été l'entité commerciale. Toutefois, cette entité ne pouvait être obtenue à partir de sources ni intérieures ni extérieures à Statistique Canada. Les seules entités qu'il nous était possible d'obtenir étaient essentiellement des entités juridiques. Il a donc fallu regrouper les entités juridiques pour reproduire approximativement des entités commerciales. L'unité de sondage a été définie comme un agrégat d'entités juridiques assujetties aux contraintes suivantes.

i. La définition de l'entité commerciale suppose que cette dernière englobe toutes les entités juridiques liées entre elles par des liens de propriété. La propriété est définie comme étant la possession de plus de 50% des actions avec droit de vote d'une entité légale. Le groupement d'entités juridiques par cette forme de propriété est limité à un seul niveau de propriété étrangère à l'extérieur du Canada.

ii. Il faut qu'une seule entité juridique canadienne ait la propriété de toutes les autres entités juridiques canadiennes faisant partie de l'entité commerciale. Cette condition est indispensable parce que les contacts en vue de l'établissement des profils avec l'entité commerciale pouvaient être pris au Canada seulement.

L'étape suivante a été de déterminer quelles unités de sondage composeraient la base de sondage et quelles données il fallait pour chacune. La base dont les entités commerciales seraient tirées pour un premier contact en vue de l'établissement d'un profil initial et à partir de laquelle on pourrait produire une première image de l'entité commerciale devait contenir toutes les entités commerciales pouvant être contactées.

Les entités peuvent être contactées en vue de l'établissement d'un profil si elles remplissent les conditions requises pour faire partie de la partie intégrée de la BDRC. Les critères déterminant si une entité commerciale doit ou non être incluse dans la partie intégrée de la BDRC sont appliqués à la structure juridique, qui décrit comment l'entité commerciale est organisée juridiquement.

Les entités juridiques peuvent faire partie de la partie intégrée de deux façons. Premièrement, si l'entité commerciale se compose d'une seule entité juridique, celle-ci entrera dans la partie intégrée si son revenu au cours de l'exercice financier considéré est supérieur à une valeur prédéterminée. Cette limite dépend de l'activité principale de l'entité juridique et de l'endroit où se trouve son siège social. Deuxièmement, si la structure juridique comprend plus d'une entité juridique, les entités juridiques feront toutes partie globalement de la partie intégrée si au moins une d'entre elles a un revenu supérieur à la limite établie pour ce genre d'entité.

Aussi, pour déterminer quelles entités commerciales pouvaient être contactées, il a fallu recueillir les renseignements suivants pour chaque entité juridique.

i. Les relations de propriété entre les entités juridiques.

ii. Le revenu de l'exercice financier considéré, l'activité principale et l'endroit où se trouve le siège social.

Pour ce qui est des entités commerciales qui remplissaient les conditions requises pour faire partie de la base de sondage et donc pour être contactées en vue de l'établissement d'un profil initial, il fallait des renseignements pour pouvoir les choisir et les contacter. Pour les choisir, il fallait les renseignements suivants:

i. La liste de toutes les industries dans lesquelles l'entité commerciale avait des activités, pour pouvoir contacter d'abord les entités du secteur du commerce de gros et/ou du commerce de détail. Les enquêtes auprès de ces entités ont nécessité, avant les autres enquêtes, un ensemble d'entités statistiques produit à partir de contacts en vue de l'établissement de profils.

ii. Le nombre de locaux d'affaires de toutes les entités commerciales comprenant une seule entité juridique ou deux entités juridiques si le propriétaire est un étranger. Cet élément d'information a déterminé le type de contact à prendre en vue de l'établissement de profils, qui pouvait être une interview téléphonique réalisée par le personnel d'un bureau régional ou une interview sur place réalisée par le personnel du bureau central ou d'un bureau régional.

iii. La province où siège la société canadienne propriétaire principale. La province a été utilisée pour répartir entre les bureaux régionaux selon leur capacité la charge de travail en termes des contacts à prendre en vue de l'établissement des profils.

Pour contacter les entités commerciales, il fallait le nom et l'adresse de l'entité juridique à la tête (excluant les propriétaires étrangers) de l'entité commerciale. Il était souhaitable de disposer en plus des données de contact et des données sur toute modalité de déclaration spéciale ayant été utilisées dans des enquêtes récentes.

2.2 Sources de données

Les sources de données qui pouvaient être utilisées étaient limitées principalement par le champ des unités que devait couvrir la base de sondage. Cette condition éliminait les listes d'éléments d'échantillons et beaucoup de listes propres à des industries précises comme des

bases de sondage. Les seules sources de données qui pouvaient être considérées étaient des listes de toutes les entités juridiques susceptibles d'entrer dans le champ des unités qui pouvaient être contactées en vue de l'établissement d'un profil et qui pouvaient produire au moins en partie certains des éléments d'information requis. Les sources de données retenues qui pouvaient être intégrées au moins partiellement par informatique étaient les suivantes:

i. La *Base de données sur les liens de parenté entre firmes* (LPF), qui est une liste de toutes les entités juridiques ayant des activités au Canada et appartenant à une entité juridique étrangère ou canadienne, et ses propriétaires. Les entités juridiques étrangères sont couvertes jusqu'au niveau qu'il faut pour remonter jusqu'au propriétaire qui en dernière analyse possède le contrôle.

ii. Le *Registre des entreprises* (RE) actuel, qui est principalement une liste de toutes les entités juridiques qui sont des employeurs. C'est dans ce registre qu'on peut trouver le nombre de locaux d'affaires d'une entité juridique, les données de contact (adresse et modalités de déclaration) utilisées pour les enquêtes et les industries dans lesquelles l'entité juridique est active.

iii. La *Base des unités de l'impôt sur le revenu des sociétés* (IRS), qui est une liste de toutes les entités juridiques qui ont produit une déclaration d'impôt sur le revenu des sociétés auprès de Revenu Canada-Impôt au cours d'une année donnée. L'activité principale, l'endroit où se trouve le siège social et le revenu de l'exercice financier sont indiqués dans cette source de données.

iv. La *Base des unités de l'impôt sur le revenu des particuliers* (IRP), qui est une liste de tous les particuliers qui ont produit une déclaration d'impôt sur le revenu auprès de Revenu Canada-Impôt au cours d'une année donnée. Les particuliers qui déclarent un revenu provenant d'un travail autonome sur leur déclaration sont des entités juridiques au sens des enquêtes économiques effectuées par Statistique Canada. On peut obtenir de cette base des données sur l'activité principale et des données de contact pour chaque particulier qui déclare un revenu provenant d'un travail autonome.

Les deux sources de données de l'impôt sur le revenu (IRS et IRP) sont des fichiers de données administratives. Les données administratives reçues tous les mois de Revenu Canada-Impôt concernant les retenues sur la paye d'un employeur sont utilisées pour mettre à jour le RE. La source de données LPF est un fichier de réponses à une enquête par recensement.

Aucune de ces sources de données ne couvre complètement l'univers des unités ni ne fournit tous les éléments d'information requis. La seule façon de couvrir tout l'univers est de les combiner. Cela vaut aussi pour certains éléments d'information requis tandis que pour les autres éléments, plus d'une source peuvent les fournir. La stratégie utilisée pour combiner ces sources de données afin d'obtenir le meilleur taux de couverture possible et la meilleure qualité des données possible est présentée dans la prochaine partie.

Une cinquième source, l'enquête sur le relevé trimestriel des états financiers, a fourni de l'information sur les entités juridiques qui préparent des états financiés consolidés. On a utiilisé cette source pour manuellement affiner les entités commerciales sur la base de sondage.

2.3 Procédures de création de la base de sondage

Le problème que pose la création de la base de sondage servant aux contacts à prendre en vue de l'établissement des profils initiaux est d'intégrer quatre sources de données qui n'ont pas été conçues pour les mêmes objectifs et qui n'ont jamais été intégrées dans cette mesure auparavant. Cette difficulté est commune à tous les utilisateurs de données administratives. La tâche était encore compliquée par le fait que c'était la première fois qu'on appliquait en même temps autant de concepts établis pour la BDRC.

Des contraintes de temps et de ressources ont obligé l'équipe chargée du projet à faire certaines hypothèses au moment de créer la base de sondage. Les hypothèses formulées étaient

toutefois justifiables puisqu'on pourrait par la suite corriger la base de sondage à l'aide des données obtenues du processus d'établissement des profils. La façon dont on a procédé pour créer la base de sondage est décrite en termes simples dans les paragraphes qui suivent.

La création de la base de sondage comportait trois étapes que nous allons décrire plus en détail un peu plus loin.

i. Construire une liste de toutes les unités pouvant éventuellement faire partie de la base de sondage.
ii. Déterminer quelles unités remplissaient les conditions requises pour faire partie de la base de sondage.
iii. Obtenir des données de sélection et de contact.

2.3.1 Création des unités susceptibles de faire partie de la base de sondage

Les unités de sondage ont été construites en groupant les entités juridiques pour créer des entités commerciales de la façon suivante. Les entités juridiques ont d'abord été groupées en structures juridiques. Une structure juridique se composait de l'ensemble des entités juridiques liées entre elles par la propriété de plus de 50 % des titres de propriété. Les relations comportant des entités juridiques étrangères étaient acceptées seulement si l'entité juridique étrangère possédait une entité juridique canadienne ou appartenait à une entité juridique canadienne. Dans les cas où une entité étrangère possédait plus d'une entité canadienne, la structure juridique a été divisée en autant d'entités commerciales qu'il y avait d'entités canadiennes appartenant directement à l'entité étrangère. Ainsi, un contact en vue de l'établissement d'un profil pouvait être pris avec le propriétaire canadien qui possède le contrôle de chaque entité commerciale résultante. Des exemples de cette façon de procéder sont donnés dans la Figure 1.

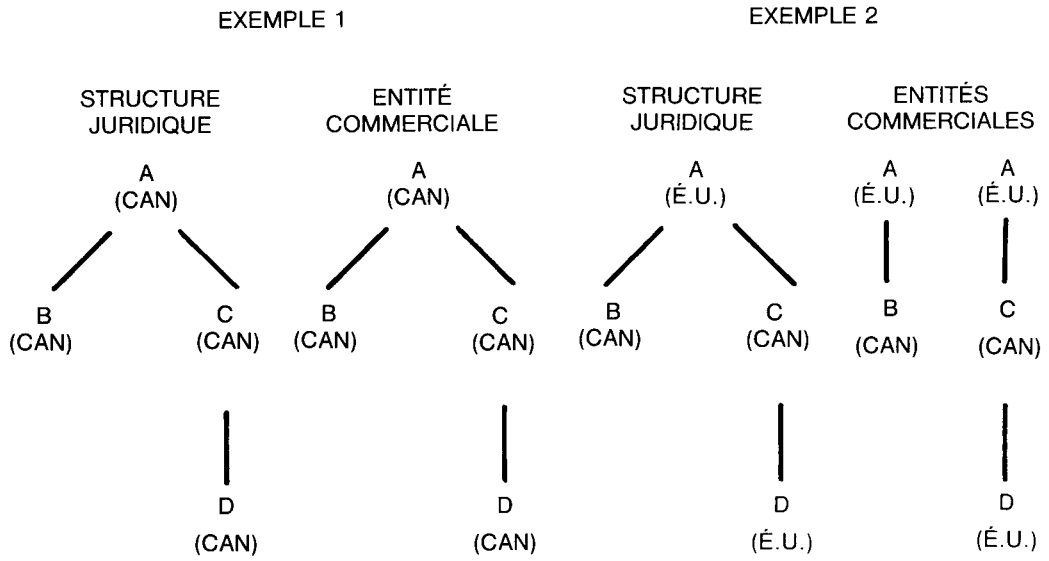

Figure 1. Définition des entités commerciales

Les particuliers ayant déclaré un revenu provenant d'un travail autonome étaient considérés comme une structure juridique contenant une seule entité juridique. Pour la construction des entités commerciales, on n'a pas tenu compte du fait que des sociétés puissent appartenir à des particuliers, ni des relations de coentreprise entre des sociétés en coparticipation.

Par conséquent, on peut considérer que l'ensemble des entités commerciales faisant partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial est formé de deux sous-ensembles mutuellement exclusifs. Le premier groupe est formé des entités juridiques qui représentent des particuliers ayant déclaré un revenu provenant d'un travail autonome. La base des unités de l'impôt sur le revenu des particuliers (IRP) contient une liste de toutes les unités de ce groupe pouvant éventuellement faire partie de la base de sondage.

Le deuxième groupe se compose des entités juridiques qui représentent les sociétés ayant des activités au Canada. La source de données sur les liens de parenté entre firmes (LPF) a été traitée de manière à fournir une liste des sociétés appartenant aux structures juridiques contenant plus d'une entité juridique. On a dressé une liste de toutes les entités juridiques n'appartenant pas à une autre entité juridique en éliminant de la base des unités de l'impôt sur le revenu des sociétés les entités juridiques qui appartenaient à une autre entité juridique ou qui étaient elles-mêmes propriétaires d'autres entités. Pour cela, il a fallu comparer les données de la base de données LPF aux données de la base IRS pour déterminer dans quelle mesure les deux bases se recoupaient. De cette façon, on a pu déterminer quelles entités juridiques figuraient dans les deux bases pour faire en sorte qu'elles n'apparaissent qu'une fois dans la base de sondage. Le couplage des entités des deux sources n'a pas été facile à effectuer et comportait une partie manuelle parce que souvent il n'existait pas de numéro d'identification commun aux enregistrements des deux sources.

2.3.2 Détermination des unités susceptibles de faire partie de la base de sondage

Les données dont on avait besoin pour déterminer si les particuliers ayant déclaré un revenu provenant d'un travail autonome pouvaient faire partie de la base de sondage figuraient dans la base IRP. Il a été très simple de déterminer si le revenu d'une entité juridique était supérieur ou non au seuil fixé au préalable pour cette entité.

La situation était plus compliquée pour les sociétés. Le couplage des données de la base de données LPF et des données de la base IRS nous a fourni les données dont nous avions besoin pour appliquer la règle du seuil limite. Toutefois, environ 20% des sociétés figurant dans la base de données LPF n'ont pu être liées à des sociétés de la base IRS. Dans ce cas, il a fallu faire une hypothèse qui a entraîné une surestimation du nombre d'entités commerciales pouvant faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. L'hypothèse était que les structures juridiques qui contenaient au moins une société non appariée satisfaisaient aux conditions d'inclusion dans la base de sondage. Sinon, les structures juridiques et, par conséquent, les entités juridiques qu'elles contenaient faisaient automatiquement partie de la base de sondage si au moins une de leurs sociétés satisfaisait à la règle du seuil limite.

2.3.3 Acquisition de données de sélection et de données de contact

L'étape précédente a abouti à une liste par approximation de toutes les entités commerciales pouvant éventuellement faire partie du champ des unités susceptibles d'être contactées en vue de l'établissement d'un profil initial. Les données de sélection et de contact décrites dans la partie 2.1 qui n'étaient pas déjà dans la base de sondage étaient disponibles du RE. La base de sondage et le RE se recoupent en partie parce que la majorité des unités de la partie de la base de sondage représentant les sociétés et une plus faible proportion des unités de la partie de la base de sondage représentant les particuliers étaient des employeurs. Il a fallu apparier les données de la base de sondage et les données du RE pour pouvoir ajouter à la base de sondage les données tirées du RE quand on constatait que des unités figuraient dans les deux sources. Autrement dit, il a fallu trouver quelles unités figuraient dans les deux sources.

Il a été encore plus difficile d'apparier ces deux sources de données que les données de la base de données LPF et les données de la base IRS. Cela était attribuable non seulement au fait qu'on ne retrouvait souvent pas de numéros d'identification communs dans les deux sources, comme dans le cas LPF-IRS, mais aussi au fait que le RE ressemble plus à la structure opérationnelle des entités commerciales qu'à leur structure juridique. Le nom et l'adresse tirés du RE ont été utilisés pour apparier les enregistrements quand il n'y a avait pas de numéro d'identification commun. Toutefois, les noms et les adresses figurant dans le RE s'appliquent souvent aux locaux «commerciaux» ou «exploitants» qui ont parfois des noms et des adresses différents des noms et des adresses «juridiques» figurant dans la base de données LPF et la base des unités IRS. Quand cela se produit, il est difficile d'établir un lien et donc d'éliminer les doubles comptes.

La base de sondage contenait certaines unités pour lesquelles on n'a pas établi de lien avec le RE soit parce que ces unités n'étaient pas des employeurs, et donc ne figuraient pas dans le RE, soit parce que les procédures de couplage n'ont pas permis d'établir le lien. Pour ces cas-là, les étapes ultérieures du processus d'établissement des profils initiaux ont été modifiées de manière à tenir compte des conditions imposées par la base de sondage. Des données de contact moins bonnes ont été tirées de la base des unités de l'impôt. Les critères de sélection ont également été modifiés de manière à tenir compte du fait qu'il puisse ne pas y avoir de données sur la répartition industrielle et les locaux d'affaires de ces entités juridiques.

Quand une entité juridique était active dans plus d'une industrie, l'activité principale pouvait être déterminée à partir des bases des unités de l'impôt sur le revenu et du RE. Il a fallu alors concilier les éléments d'information communs quand ils différaient. Dans ce cas, l'industrie indiquée dans le RE a été utilisée étant donné que l'on jugeait cette source de renseignements plus fiable.

La figure 2 illustre par une figure (qui n'est pas à l'échelle) la base de sondage ainsi créée.

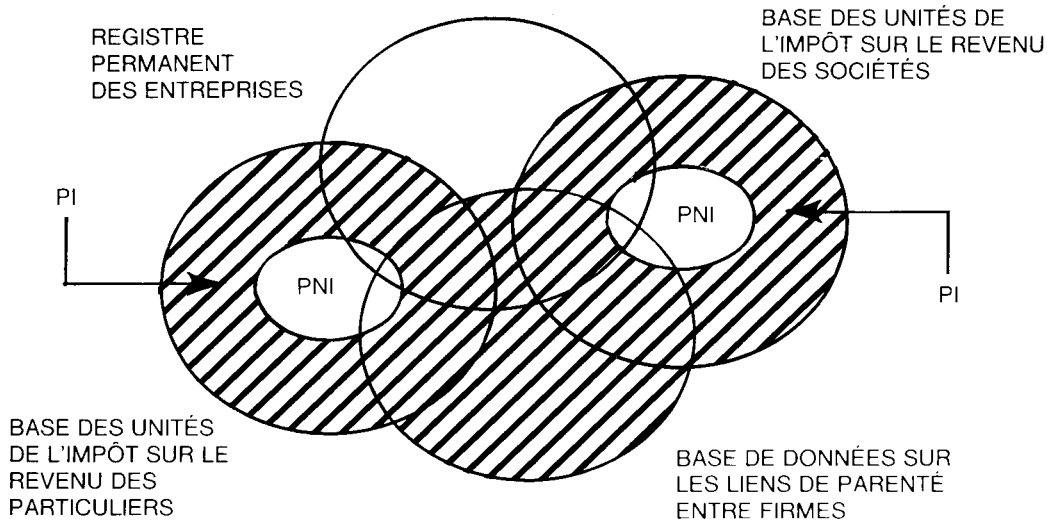

Figure 2. Base de sondage en vue de l'établissement des profils initiaux

2.3.4 Évaluation de la base de sondage

La qualité de la base de sondage obtenue a été évaluée de trois façons. On a d'abord vérifié que la base de sondage concordait avec les spécifications développées pour sa création.

Deuxièmement, on a comparé diverses distributions des entités juridiques figurant dans la base de sondage avec les mêmes distributions produites de façon indépendante à partir d'une simulation de la partie intégrée. Les distributions ne différaient pas de façon significative.

Enfin, on a évalué la base de sondage en la comparant au RE. Un échantillon de 30 des plus grandes unités du RE a été apparié à la base de sondage créée en vue de l'établissement des profils initiaux. On a retrouvé toutes les unités, quoique très difficilement, parce que les deux sources n'utilisent pas les mêmes concepts.

2.4 Conclusion

La stratégie élaborée pour créer la base de sondage qu'on vient de décrire reposait sur des hypothèses simples au sujet de la couverture, de la qualité des données et de la façon dont fonctionnent les entités commerciales. On a souvent utilisé des «raccourcis» pour remplir les conditions requises par la base de sondage. On a estimé que cette approche était parfaitement justifiée étant donné qu'on jugeait que le rôle de la base de sondage était de donner des entités commerciales une première image qui pouvait ensuite être mise à jour pendant le processus d'établissement des profils. Les conséquences de ces hypothèses sont analysées dans les paragraphes qui suivent.

La population des entités commerciales pouvant faire partie du champ des entités susceptibles d'être contactées en vue de l'établissement d'un profil initial peut contenir des unités comptées deux fois et des unités hors du champ. Si c'est le cas, on prendra plus de contacts qu'il n'est nécessaire. Cela augmentera les coûts de production de Statistique Canada et accroîtra indûment le fardeau de réponse de certains répondants en les obligeant à répondre deux fois aux mêmes questions. Enfin, l'image de Statistique Canada pourrait être ternie un peu.

La population pourrait être sous-estimée. Dans ce cas-là, on pourra établir le profil des unités manquantes un peu plus tard. Cela risque de retarder l'introduction de nouvelles grandes unités dans la partie intégrée de la BDRC. Entre-temps, les unités manquantes seraient couvertes par la partie non intégrée et non par la partie intégrée.

L'imprécision des données de sélection et/ou de contact pourrait compliquer ou retarder les contacts jusqu'à ce qu'on obtienne des données précises. Dans ces cas-là, il en résulte également que la BDRC sera imprécise jusqu'à ce que tous les profils aient été achevés.

Ces divers résultats montrent les complications que peut entraîner l'utilisation de données administratives. Ils montrent également qu'il faut bien s'assurer que les données administratives utilisées concordent avec les objectifs fixés. Des exemples ont été donnés des types de compromis qu'il faut faire quand on ne peut raisonablement pas obtenir une telle compatibilité.

3. UTILISATION DE DONNÉES ADMINISTRATIVES POUR L'ÉTABLISSEMENT DES PROFILS ULTÉRIEURS

3.1 Établissement cyclique et établissement ponctuel de profils

Il y aura deux types de mise à jour de profils initiaux, à savoir l'établissement cyclique de profils et l'établissement ponctuel de profils. Chaque type est expliqué ci-dessous.

L'établissement cyclique de profils est le processus par lequel on s'assure que le profil de toutes les entités commerciales de la population est refait au bout d'un certain temps. Compte tenu des prévisions budgétaires actuelles, on prévoit que cette période sera de deux ans. Le temps écoulé depuis le dernier profil sera le premier critère utilisé pour déterminer si cette entité doit être soumise au processus d'établissement cyclique des profils. On tiendra compte d'autres facteurs pour classer par ordre de priorité les entités dont il faut refaire le profil.

L'établissement ponctuel de profils est le processus d'établissement de profils des entités commerciales qui est déclenché parce qu'on a reçu d'une autre source des renseignements selon lesquels des changements se seraient produits dans l'entité commerciale et que l'image statistique de l'entité commerciale dans le registre peut n'être plus valide. Le processus d'établissement ponctuel de profils fera que la BDRC sera plus à jour que si l'on utilisait seulement le mécanisme d'établissement cyclique des profils. Les divers fichiers de données administratives reçus régulièrement à Statistique Canada font partie des autres sources d'information sur les changements.

3.2 Sources de données administratives pouvant être utilisées

Les trois sources de données administratives que Statistique Canada peut utiliser pour mettre à jour son registre central et dont nous parlons dans le présent document sont:

- la Base des unités de l'impôt sur le revenu des particuliers,
- la Base des unités de l'impôt sur le revenu des sociétés, et
- les données sur les comptes de retenue sur la paye saisies par les administrations fiscales.

En général, les particuliers et les sociétés produisent une seule déclaration d'impôt pour une année de référence donnée. Il se peut toutefois qu'ils en produisent plus d'une parce que, par exemple, une société aurait changé la fin de son exercice financier avec l'approbation des administrations fiscales. On peut quand même dire que les déclarations d'impôt sur le revenu sont une source de données annuelles sur les changements.

Ce n'est pas seulement une fois par année que les bases des unités de l'impôt sur le revenu sont communiquées à Statistique Canada. En fait, Statistique Canada reçoit régulièrement pendant deux ans les fichiers de données fiscales se rapportant à une année de référence donnée. Par conséquent, on pourrait mettre à jour tous les mois le registre à partir des données fiscales, mais chaque enregistrement du registre serait généralement mis à jour une fois par année seulement.

Par ailleurs, on s'attend qu'en règle générale les employeurs s'acquittent de leurs comptes de retenue sur la paye une fois par mois. Ainsi, Statistique Canada reçoit un fichier de données des comptes de retenue sur la paye une fois par mois. Par conséquent, on peut mettre à jour le registre tous les mois à partir des données des comptes de retenue sur la paye, et chaque enregistrement du registre peut en théorie être modifié tous les mois.

Il convient de noter que d'autres sources de données administratives peuvent aussi être utilisées. On n'en parle pas dans le présent document parce qu'elles ne couvrent pas toute la population, ni sur une base régulière. Il convient toutefois de les citer. Ce sont:

- l'information limitée recueillie par les administrations fiscales sur les sociétés qui n'ont pas produit de déclaration d'impôt mais dont on pense qu'elles sont actives,
- d'autres données recueillies d'un échantillon de déclarations d'impôt sur le revenu par Statistique Canada, et
- les données recueillies par Statistique Canada à même les formules fiscales remplies par les employeurs demandant qu'on leur ouvre un compte de retenue sur la paye.

3.3 Signaux de changement

On a mis au point un système de signaux de changement qui proviendraient de chacune des sources de données administratives décrites dans la partie précédente. Ces signaux, rapportés à des enregistrements administratifs, indiquent quelles entités statistiques ont pu subir des changements. Ils indiquent également aux responsables du registre qu'il peut- être souhaitable de refaire le profil de ces entités statistiques à l'aide du processus d'établissement ponctuel de profils pour tenir le registre à jour.

Les signaux dépendent de la source administrative. Dans chacune des trois sources énumérées dans la partie 3.2, les signaux sont des tests de comparaison entre les nouvelles données reçues pour un enregistrement donné et les données précédentes reçues de la même source pour l'enregistrement en question. Les tests peuvent porter sur une seule zone ou sur un groupe de zones et être conditionnel à une ou plus d'une zone. Ces tests de comparaison tentent de refléter les événements du monde réel qui influent sur les entités statistiques et non pas seulement ceux qui influent sur les entités administratives. Il faut se rappeler que les entités statistiques ont été créées aux fins des programmes de statistiques économiques et souvent diffèrent complètement des entités juridiques ou administratives. Aussi, ces tests de comparaison devraient permettre le plus possible de trouver les changements survenus dans les données administratives qui reflètent un changement dans les entités statistiques. Par exemple, un changement de propriétaire d'une usine de fabrication peut signifier qu'un enregistrement administratif disparaîtra et qu'un autre apparaîtra. Cela peut par contre n'avoir aucun effet sur les entités statistiques étant donné que le même établissement avec le même degré de capacité de fournir les données requises peut continuer d'exister.

Si la base de sondage était mise à jour directement à partir des changements observés dans les enregistrements administratifs sans aucun autre contrôle, il y aurait une forte proportion de créations et de disparitions apparentes d'entités et un risque proportionnel de couverture incomplète ou de couverture en double. C'est pourquoi il faut aussi contacter les répondants ou au moins faire des recherches internes à l'aide de tous les documents qu'on peut obtenir pour trouver ce qui est arrivé aux entités statistiques dans le cas des enregistrements administratifs signalés. Le processus de «traduction» n'est pas ce qu'il y a de plus simple et la recherche d'une solution constitue justement l'objectif du processus d'établissement ponctuel de profils.

Le nombre de signaux retenus pour chaque source et quelques exemples de signaux sont présentés dans le tableau 1. Il faut toutefois noter les points suivants quand on regarde les données sur le nombre de signaux. Certains signaux sont très raffinés, d'autres pas. Dans bien des cas, on a décidé de diviser le signal initial en deux sous-signaux mutuellement exclusifs parce qu'on estimait que cela pouvait contribuer davantage à déterminer les bonnes mesures à prendre à partir du signal. L'exemple le plus simple concerne les comptes de retenue sur la paye. Dix-huit des quarante signaux représentent des changements dans le nombre estimatif d'employés couverts par le compte. Les 18 signaux permettent de distinguer entre des augmentations et des diminutions du nombre estimatif et tiennent compte de l'ampleur des augmentations et des diminutions. On a estimé que cette façon de procéder aiderait à classer par ordre de priorité les travaux manuels à accomplir. On pourrait néanmoins considérer tous ces signaux comme un seul signal.

Tableau 1

Signaux par source administrative

Source administrative	Nombre de signaux distincts	Exemples
Déclarations annuelles d'impôt sur le revenu des particuliers	50	Passage d'une seule à plus d'une province d'imposition
Déclarations annuelles d'impôt sur le revenu des sociétés	49	Début d'une coentreprise
Comptes mensuels de retenue sur la paye	38	Nouveau compte avec la description du nom qui identifie une société

Même si les déclarations d'impôt sur le revenu sont dépouillées régulièrement, on s'attend qu'une déclaration donnée produise en général des signaux au maximum une fois par année de référence, tandis qu'un compte de retenue sur la paye pourrait produire un ou plus d'un signal tous les mois. Mais ce qui nous intéresse le plus, ce n'est pas le nombre de signaux que peut transmettre une source, mais le nombre d'enregistrements que ces signaux permettent de trouver. Cela donnerait une idée des ressources en personnel de bureau à investir pour mettre à jour le registre à partir de sources administratives. On a donc fait une étude de simulation pour répondre à cette question.

3.4 Étude de simulation

L'étude de simulation a consisté à appliquer les signaux décrits précédemment aux populations suivantes:

- les déclarations d'impôt sur le revenu des particuliers pour les exercices financiers prenant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les déclarations d'impôt sur le revenu des sociétés pour les exercices financiers prenant fin en 1984, pour détecter les changements survenus durant ces exercices financiers;
- les comptes de retenue sur la paye du début d'octobre 1985, pour détecter les changements survenus depuis le début de septembre 1985.

Les résultats de l'étude de simulation sont présentés dans le tableau 2. On peut faire les observations suivantes sur les résultats obtenus.

- Beaucoup de déclarations d'impôt sur le revenu produisent des signaux: seulement un huitième environ des déclarations d'impôt sur le revenu des particuliers et un cinquième des déclarations d'impôt sur le revenu des sociétés n'ont pas produit de signal.
- Durant la période d'un mois observée, 8,258 comptes de retenue sur la paye ont produit un signal pour le mois considéré. Si l'on suppose que les signaux émis par les comptes de retenue sur la paye sont distribués uniformément d'un mois à l'autre, cela ferait presque 100,000 comptes signalés par année. Il convient toutefois de remarquer qu'il est fort possible qu'un même compte soit signalé plus d'un mois et qu'on risque donc de compter au moins deux fois les signaux si on les additionne sur une période d'un an.
- Si l'on additionne tous les enregistrements signalés au cours de l'année, cela donne au total 244,269 enregistrements signalés. Il est toutefois évident qu'un même signal peut se trouver dans plus d'une source administrative. Par exemple, un changement de raison sociale d'une entreprise peut être indiqué dans la déclaration d'impôt produite par l'entreprise et dans chacun de ces deux comptes de retenue sur la paye.

Tableau 2

Résultats de l'étude de simulation

Source administrative	Population totale	Nombre d'enregistrements signalés	Nombre d'enregistrements signalés en pourcentage de la population totale
Déclarations d'impôt sur le revenu des particuliers	72,190	63,446	87.9
Déclarations d'impôt sur le sur le revenu des sociétés	102,688	81,727	79.6
Comptes de retenue sur la paye	134,973	8,258	6.1

3.5 Questions soulevées

Les résultats de l'étude de simulation et l'analyse du rôle des signaux soulèvent un certain nombre de questions au sujet de l'établissement des profils.

Six de ces questions sont présentées ci-dessous.

3.5.1 Dans quelle mesure les signaux permettent de détecter les changements survenus dans les entités statistiques

Les signaux indiqueront plus ou moins fidèlement quelles entités juridiques et/ou exploitantes ont subi des changements réels qui ont un effet sur les entités statistiques. Il faudra ensuite mettre à jour le registre central pour maintenir la qualité des produits statistiques. Les signaux reflètent-ils réellement les événements du monde réel qui influent sur les entités statistiques ou est-ce qu'il y en a qui ne correspondent à aucun effet? Si certains n'ont pas de signification, cela va engendrer du travail inutile.

Une enquête à petite échelle a été menée en 1986 pour déterminer dans quelle mesure les signaux permettaient de déceler les changements survenus dans les entités statistiques. Toutefois, pour diverses raisons, les seuls signaux qu'on a pu utiliser étaient ceux de l'étude de simulation. Ces signaux ont trait aux changements survenus dans les déclarations d'impôt entre les années d'imposition 1983 et 1984. Mais le décalage entre la période de référence des signaux et la période d'enquête (1986) a posé des problèmes de mémoire aux répondants. C'est ainsi que les répondants ont inclus des événements qui ont eu lieu après la période de référence ou ont oublié des événements qui ont eu lieu pendant la période de référence. On a donc jugé que l'enquête n'était pas concluante, et aucun autre essai n'a été tenté depuis ce temps-là.

3.5.2 Répétition de signaux

Les signaux seront reçus à plusieurs moments et de sources différentes indépendantes les unes des autres. Les déclarations d'impôt sur le revenu, en particulier, entraînent des délais sensibles. Au moment où un signal est reçu, la BDRC pourra déjà avoir été mise à jour pour tenir compte de l'événement du monde réel indiqué par ce signal. Cette mise à jour pourra avoir été faite à l'occasion du traitement d'un signal venu d'une autre source ou du processus cyclique d'établissement des profils ou encore de l'incorporation de renseignements tirés d'autres enquêtes. Par conséquent, on ne peut pas traiter les signaux indépendamment de la BDRC pour décider de procéder à l'établissement ponctuel d'un profil. Toutefois, comment devrait-on vérifier un signal en se référant à la BDRC pour voir si la BDRC a déjà été mise à jour? Par exemple, dans le cas où une déclaration d'impôt sur le revenu des sociétés indique une grosse augmentation de revenu, comment peut-on vérifier que la BDRC a déjà été mise à jour pour tenir compte de l'événement du monde réel sous-jacent à cette augmentation quand on ne sait même pas quel est cet événement?

3.5.3 Omission de signaux de changement

De même, certains enregistrements ne seront pas signalés. Est-ce que l'absence de signaux signifiera nécessairement qu'aucun événement du monde réel ne s'est produit qui oblige à mettre à jour la structure statistique? Devrait-on mettre au point d'autres signaux pour couvrir les omissions? Encore ici, l'étude mentionnée précédemment n'a pas donné de réponses concluantes à ces questions.

3.5.4 Possibilité d'obtenir des ressources en quantité suffisante pour traiter les enregistrements signalés

Comme l'étude de simulation l'a montré, beaucoup d'enregistrements seront signalés qu'il faudra traiter en partie à la main. Il est probable qu'il n'y aura pas assez de ressources pour accomplir tout ce travail. Comment devrait-on procéder pour déterminer quelle proportion de l'ensemble des ressources devrait être consacrée à l'établissement ponctuel de profils et pour déterminer la façon d'utiliser les ressources pour traiter les enregistrements signalés? Si une rareté de ressources oblige à ignorer un certain nombre de signaux, comment allons-nous déterminer quels signaux il faut laisser tomber?

3.5.5 Fardeau de réponse

Les résultats de l'étude de simulation semblent indiquer qu'il faudra contacter les entreprises plus souvent qu'une fois tous les deux ans si l'on veut vérifier les changements à apporter à la base de sondage par d'autres moyens que les enquêtes régulières. Cela augmentera nécessairement le fardeau de réponse. Peut-on trouver un compromis entre l'obligation d'augmenter le fardeau de réponse pour tenir le registre à jour et la nécessité d'empêcher que le registre ne cesse d'être à jour? Quel devrait être ce compromis?

3.5.6 Rôle du processus d'établissement cyclique des profils

Le grand nombre d'enregistrements signalés par les déclarations d'impôt dans l'étude de simulation amène à s'interroger sur l'utilité du processus d'établissement cyclique des profils. On peut déduire des résultats de cette étude que très peu d'enregistrements devront être soumis au processus cyclique plutôt qu'au processus ponctuel. Supposons par exemple que les résultats de l'étude de simulation concernant le nombre d'enregistrements signalés continuent d'être vrais une deuxième année. Supposons ensuite que les enregistrements signalés la deuxième année ne sont pas tous les mêmes que les enregistrements signalés la première année, mais qu'un certain nombre de nouveaux enregistrements sont signalés et que certains enregistrements de la dernière année ne sont pas signalés la deuxième année. On peut alors supposer sans grand risque d'erreur que très peu d'enregistrements ne recevront aucun signal pendant deux ans. Il restera très peu d'enregistrements qui n'auront pas été signalés l'une ou l'autre année. Ce nombre représentera en fait le nombre maximum d'unités de la population à soumettre au processus d'établissement cyclique des profils. Faudra-t-il établir absolument le profil de ces entités, sachant qu'elles ne sont signalées ni par les comptes de retenue sur la paye, ni par les déclarations d'impôt?

4. CONCLUSION

Dans la partie 2, nous avons montré comment les données administratives ont été utilisées pour construire une base de sondage en vue de l'établissement de profils initiaux. Les données administratives offraient un taux de couverture très élevé. Toutefois, nous avons aussi vu que les différences conceptuelles existant entre nos besoins et les données administratives amenaient des complications qui nous ont obligés à faire des hypothèses et des compromis simplificateurs.

La base de sondage que nous avons créée incluait les éléments nécessaires à l'établissement des profils initiaux de toutes les entités commerciales à l'exception des plus complexes. Dans ces derniers cas, on ne pouvait pas accepter les approximations données par la base de sondage. Il a donc fallu faire beaucoup de recherches sur ces entités commerciales en utilisant d'autres sources d'information comme des rapports annuels publics et des réponses à des enquêtes.

La base de sondage a également joué un rôle important dans la construction et la mise en vigueur de la BDRC. Elle a servi avec le Registre des entreprises à déterminer quelles entités seraient incluses dans la partie intégrée.

La méthode suivant laquelle les données administratives seront utilisées pour établir les profils de mise à jour a été décrite dans la partie 3. Des signaux de changement seront tirés de diverses sources de données administratives et produiront des demandes de vérification des profils. Beaucoup de questions ont été soulevées à cet égard. Les diverses équipes chargées de mettre en pratique la stratégie de mise à jour de la BDRC s'emploient actuellement à trouver des solutions à ces questions. Une des solutions à une partie des problèmes serait de classer les signaux par ordre de priorité selon, par exemple, la durée de la période depuis la dernière fois qu'on a établi le profil d'une entité. Une autre solution serait de mettre au point un système autodidacte. L'expérience dira quels signaux sont utiles et devront être conservés. Il reste donc encore beaucoup de travail à faire avant que le processus soit tout à fait opérationnel.

BIBLIOGRAPHIE

COLLEDGE, M., et LUSSIER, R. (1985). A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 123-131.

COLLEDGE, M. (1987). Projet de remaniement des enquêtes-entreprises - Mise en pratique d'une nouvelle stratégie à Statistique Canada. Document présenté à la *Bureau of the Census Third Annual Research Conference*.

STATISTIQUE CANADA (1987). Version 4.2 of the CFDB Data Dictionary. Projet de remaniement des enquêtes-entreprises, document de travail, 4 mars 1987.