Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d’enquête

L’échantillonnage pour les enquêtes-entreprises à Statistique Canada

par M.A. Hidiroglou

Date de diffusion : le 23 décembre 2025

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, décembre 2025
Vol. 51, nº 2, pp. 371-412
Statistique Canada, nº 12-001-X au catalogue

L'échantillonnage pour les enquêtes-entreprises à Statistique Canada

M.A. Hidiroglou[1]

Résumé

Le présent article examine les complexités méthodologiques associées à la conception des enquêtes-entreprises, en mettant particulièrement l'accent sur les stratégies d'échantillonnage mises en œuvre par les organismes nationaux de statistique. Il aborde les défis inhérents à la nature dynamique de la population des entreprises, qui nécessite des mises à jour constantes de la base de sondage afin d'en garantir la représentativité et la pertinence. Les considérations essentielles en matière de plan de sondage comprennent la détermination de la taille optimale des échantillons, la stratification selon des dimensions clés telles que l'industrie, la région géographique et la taille de l'entreprise, ainsi que le traitement des créations d'entreprises et l'exclusion des unités inactives (ou « unités disparues »). L'article applique la méthode de répartition de puissance de Bankier (1988) à un plan de stratification à deux dimensions par industrie et région géographique, puis en évalue le rendement en comparant les coefficients de variation obtenus à ceux produits par un algorithme du ratissage appliqué aux coefficients marginaux. L'approche est ensuite étendue à un contexte multivarié afin de couvrir plusieurs domaines d'estimation. La discussion englobe aussi des enjeux pratiques liés au renouvellement et à la coordination des échantillons, essentiels pour maintenir la qualité des données et réduire le fardeau des répondants au fil du temps.

Mots-clés : Coordination des échantillons; méthode de Bankier; microstrates; nombres aléatoires permanents; registre des entreprises; renouvellement; répartition de Neyman; répartition de puissance; retrait des unités disparues.

1. Introduction

Les enquêtes-entreprises sont généralement menées sur une base mensuelle, trimestrielle ou annuelle afin de produire des estimations de totaux, de moyennes, de ratios et d'évolutions temporelles pour un ensemble d'indicateurs économiques et structurels clés. Les défis méthodologiques liés à la conception des enquêtes-entreprises sont largement partagés par les organismes nationaux de statistique (ONS) et s'avèrent généralement plus complexes que ceux rencontrés dans les enquêtes-ménages ou les enquêtes sociales. Cela s'explique par la nature hautement dynamique et hétérogène de la population des entreprises.

Ces enquêtes sont des outils essentiels pour recueillir des données auprès des entreprises à des fins statistiques et stratégiques variées. Les ONS les utilisent pour compiler des indicateurs macroéconomiques essentiels, tels que le produit intérieur brut (PIB), les investissements, les dépenses publiques et les statistiques du commerce extérieur pour diverses périodes de référence. L'étendue des données recueillies par ces enquêtes-entreprises est vaste. Elles peuvent englober les domaines comme la production (extrants, intrants, transport, émissions environnementales), les ventes (commerce de gros et de détail), les flux de produits (stocks, expéditions, nouvelles commandes), la performance financière (revenus, dépenses, actifs, passifs), l'emploi (niveaux d'emploi, données sur la paye, données démographiques des employés), ainsi que les prix (indices des prix à la production et à la consommation).

1. M.A. Hidiroglou, Statistique Canada alumnus, Ottawa, Ontario. Courriel : hidirog@yahoo.ca.

Compte tenu de l'hétérogénéité marquée des entreprises en matière de taille et de classification industrielle, des plans de sondage rigoureux sont indispensables. Les enquêtes-entreprises sélectionnent généralement leurs échantillons à partir des registres des entreprises, établis à partir de sources administratives et comprenant les principales coordonnées (noms de l'entreprise, adresses, renseignements sur la communication propres à l'entreprise). L'objectif du présent article est de décrire et de proposer des solutions aux problèmes d'échantillonnage dans les enquêtes-entreprises. Le choix de la méthode d'échantillonnage influe directement sur l'efficacité et la précision de l'enquête. La taille de l'échantillon joue un rôle crucial dans la fiabilité des résultats. Un échantillon trop réduit peut mener à des estimations peu fiables, tandis qu'un échantillon excessivement vaste engendre des coûts superflus et accroît inutilement le fardeau opérationnel et celui imposé aux répondants. Les concepteurs du plan de sondage doivent donc calculer soigneusement la taille d'échantillon appropriée, en tenant compte du niveau de précision visé, de la variabilité de la population cible, et en équilibrant la rigueur statistique et les contraintes pratiques. La population cible (également appelée *portée* de l'enquête) peut couvrir l'ensemble des unités d'un pays, celles d'un secteur industriel particulier, ou celles d'une région particulière.

L'article est organisé comme suit. La section 2 présente l'évolution du Registre des entreprises (RE) de Statistique Canada depuis sa création dans les années 1970, en mettant l'accent sur son architecture actuelle en lien avec les applications d'échantillonnage. La section 3 expose le cadre méthodologique de la stratification des unités statistiques recensées dans le RE selon trois principales dimensions : l'industrie, la région géographique et la taille de l'entreprise. La section 4 passe en revue les procédures de sélection initiale de l'échantillon, de renouvellement de l'échantillon et d'intégration des nouvelles entreprises (créations) dans les enquêtes-entreprises en cours. Ces méthodes visent à garantir que l'échantillon demeure représentatif d'une population en évolution au fil du temps, afin d'appuyer la production d'estimations exactes et fiables. La section 5 traite de la détermination et du retrait des unités inactives (« unités disparues ») tant dans l'échantillon qu'en dehors de l'échantillon. Un défi important provient du fait que les unités connues comme étant inactives (ou « disparues ») sont découvertes en proportion plus élevée dans l'échantillon qu'à l'extérieur. De plus, une partie de la population hors échantillon peut comprendre des entreprises inactives qui n'ont pas encore été formellement désignées comme telles (disparition inconnue d'une entreprise). Les unités disparues connues dans l'échantillon sont généralement repérées par le commentaire direct des répondants ou à l'aide de sources auxiliaires, tandis que celles situées hors échantillon ne peuvent être reconnues qu'à partir de sources indépendantes. Cette asymétrie souligne la nécessité de méthodologies solides pour gérer le retrait des unités inactives dans l'ensemble de la population observée. Un résumé est présenté à la section 6.

2. Registre des entreprises

2.1 Un peu d'histoire

Au début des années 1970, Statistique Canada a créé son Registre des entreprises (RE) pour servir de cadre fondamental pour diverses enquêtes économiques. Initialement, le RE était établi principalement à

partir des données de retenues sur la paye fournies par Revenu Canada pour les petites entreprises et à partir des mises à jour consécutives aux enquêtes pour les grandes entreprises. Cependant, avec le temps, il est apparu que le RE ne répondait pas adéquatement aux besoins changeants des programmes d'enquêtes économiques, particulièrement en ce qui concerne la couverture, l'actualité et les outils analytiques. L'une des principales lacunes était une couverture incomplète : le RE manquait de contrôle centralisé sur les données nécessaires et demeurait fragmenté entre diverses divisions et bases de données. Par conséquent, de nombreuses enquêtes majeures ont choisi de ne pas passer au RE et ont préféré maintenir leurs propres bases de sondage indépendantes, invoquant les limites du RE pour leurs besoins particuliers.

En réponse à ces problèmes, Statistique Canada a lancé en 1984 le projet de remaniement des enquêtes-entreprises. Cette vaste initiative visait à moderniser l'infrastructure de l'organisme pour les opérations d'enquête économique. L'objectif principal du projet était d'unifier et d'optimiser la maintenance des structures d'entreprises pour l'ensemble des enquêtes. À l'époque, les systèmes existants étaient disparates et inefficaces, répartis dans plusieurs divisions et bases de données. Le remaniement visait à consolider ces systèmes dans un cadre cohérent, afin d'améliorer l'uniformité des données, de faciliter l'accès à l'information sur les entreprises et d'accroître l'efficacité opérationnelle. Un point central était l'augmentation de l'utilisation de sources de données administratives, telles que les déclarations de revenus et d'autres renseignements commerciaux détenus par le gouvernement, afin de réduire la redondance et d'améliorer la qualité des données.

La nouvelle infrastructure d'enquêtes est entrée en vigueur en mars 1988, à commencer par l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD). Ces enquêtes étaient idéales pour une première mise en œuvre, car elles dépendaient d'une base de sondage des entreprises actuelle et fiable pour collecter les données sur le commerce de gros et de détail. Même si les deux enquêtes partageaient une architecture modulaire commune, comprenant des composantes pour la sélection de l'échantillon, la collecte de données, la vérification, l'imputation et l'estimation, elles différaient dans certains détails méthodologiques. Par exemple, l'Enquête sur le commerce de détail ciblait les emplacements, tandis que l'Enquête sur le commerce de gros ciblait les établissements. Elles différaient aussi par les variables recueillies, les méthodes d'imputation et les procédures d'estimation.

La conception modulaire a été spécialement conçue pour être flexible, permettant son adaptation et sa réutilisation dans un large éventail d'enquêtes-entreprises. Cela a marqué le début d'une approche généralisée des méthodes d'enquête-entreprise. Une documentation détaillée de ce cadre se trouve dans Hidiroglou (1986a). Le plan de sondage de l'EMCGD a connu plusieurs améliorations au fil du temps, comme décrit par Trépanier, Babyak, Marchand, Bissonnette et Saint-Pierre (1998), Bérard (2001), Trépanier (2004) et Ferland et Batten (2007).

La mise en œuvre réussie de l'EMCGD a marqué un tournant dans la modernisation de l'infrastructure des enquêtes-entreprises à Statistique Canada. Tablant sur cette réussite, la méthodologie et les systèmes de traitement associés ont été progressivement élargis pour inclure un éventail plus large d'entreprises, notamment celles dans le secteur manufacturier, des services et d'autres secteurs. Ce déploiement progressif

a favorisé une plus grande cohérence, une meilleure homogénéité et la rentabilité dans l'ensemble des programmes d'enquête.

Une avancée décisive a eu lieu en 1994 avec l'introduction du numéro d'entreprise (NE) à neuf chiffres par l'Agence du revenu du Canada (ARC), visant à simplifier l'identification des entreprises à des fins fiscales et réglementaires (Saint-Louis, 2008). La racine du NE se compose des neuf premiers chiffres du NE à 15 chiffres. Elle est commune à toutes les entités d'une même entreprise et permet à l'ARC de déterminer facilement les entités liées. Pour obtenir plus de renseignements sur l'attribution du NE, consultez le document de l'ARC *« Le numéro d'entreprise et vos comptes de programme de l'Agence du revenu du Canada ».*

L'intégration du NE dans le RE de Statistique Canada en 1997 a représenté une amélioration majeure de la capacité de l'organisme à lier les ensembles de données administratives. Cette intégration a accru la précision, l'exhaustivité et le niveau de détail des statistiques économiques provinciales, en permettant des liens cohérents entre diverses sources de données telles que les déclarations de revenus et les enquêtes sur la population active. L'adoption d'un identificateur d'entreprise unique a permis de réduire les problèmes liés à des identificateurs incohérents ou incomplets, améliorant ainsi la qualité et la fiabilité des données.

Parallèlement, Statistique Canada a entrepris un remaniement complet de son architecture des enquêtes-entreprises annuelles, menant à la création de l'Enquête unifiée auprès des entreprises (EUE) (Brodeur, Koumanakos, Leduc, Rancourt et Wilson, 2005). Conçue dans le cadre du vaste Projet d'amélioration des statistiques économiques provinciales (PASEP) (Statistique Canada, 1997), l'EUE visait à regrouper toutes les enquêtes-entreprises annuelles dans un seul cadre harmonisé. Ses principaux objectifs étaient d'améliorer la cohérence, l'homogénéité et l'exhaustivité des statistiques des entreprises, tant d'un point de vue provincial qu'industriel.

En 1998, Statistique Canada a entrepris une refonte majeure du RE, s'appuyant sur la précédente mise en œuvre du NE et s'harmonisant aux objectifs de l'initiative du PASEP. Cette transformation a marqué une étape importante vers la modernisation de l'infrastructure des statistiques des entreprises de l'organisme. Une mise à jour subséquente en 2005 a encore simplifié les processus et consolidé les fondements du cadre de l'EUE. D'autres améliorations ont été apportées à l'infrastructure de la statistique des entreprises en 2010 grâce au lancement du Programme intégré de la statistique des entreprises (PISE), visant à accroître l'efficacité, la cohérence et la réactivité des enquêtes-entreprises (Daoust, Mireuta et Kleim, 2023). Le PISE reposait sur six objectifs principaux : i) améliorer la qualité des données; ii) réduire le fardeau de réponse grâce à des stratégies d'échantillonnage coordonnées; iii) moderniser l'infrastructure de traitement des données; iv) intégrer la majorité des enquêtes économiques dans un cadre unifié; v) normaliser les processus et les outils dans l'ensemble des programmes; vi) accroître la faculté d'adaptation aux besoins en constante évolution en matière de données.

Un effort comparable de modernisation a été mené par le U.S. Census Bureau en 2018 (Kirkendall, White Jr, Citro et Abraham, 2018), reflétant une tendance internationale plus large vers des systèmes de collecte de données économiques plus intégrés et efficaces.

2.2 Structure du RE

Le RE est une ressource essentielle pour la réalisation des enquêtes-entreprises et la production de statistiques économiques fiables au Canada. Sa fonction principale est de servir de base de sondage de haute qualité pour les enquêtes économiques menées par Statistique Canada. Le RE comprend toutes les entités économiques participant à la production de biens et de services, y compris les entreprises constituées en société ou les entreprises non constituées en société, les établissements commerciaux, les organisations sans but lucratif, les institutions religieuses et les organismes gouvernementaux.

En termes statistiques, une entreprise est définie comme une unité économique, telle qu'un établissement ou une ferme, qui utilise des ressources (main-d'œuvre, capital et matières premières) pour produire des biens ou des services. Les entreprises opèrent dans divers secteurs économiques, y compris, mais sans s'y limiter, le commerce de détail, le commerce de gros, les services, le secteur manufacturier, la construction, l'énergie, le transport, l'agriculture et le commerce international. Une enquête-entreprises recueille, à des fins statistiques, des données issues d'un échantillon d'entreprises ou d'établissements.

Chaque entreprise du RE est décrite par une série d'attributs regroupés en cinq grandes catégories : données d'identification, données de classification, coordonnées, situation d'activité et données de mise à jour/de couplage.

- Les données d'identification caractérisent chaque unité au moyen des éléments suivants : nom de l'entreprise, adresse physique, identificateurs alphanumériques (le NE).
- Les données de classification comprennent la taille, les catégorisations industrielles et géographiques, qui sont essentielles pour la stratification de la population et la constitution d'échantillons représentatifs.
- Les coordonnées sont les renseignements nécessaires à l'administration de l'enquête comme les personnes-ressources, les adresses postales, les numéros de téléphone et l'historique des réponses aux enquêtes.
- La situation d'activité indique si l'unité est actuellement opérationnelle (active, en activité, en saison) ou inactive (disparue, hors saison).
- Les données de mise à jour et de couplage servent à suivre l'évolution temporelle des entités commerciales, notamment les changements de structure ou d'activité par des enregistrements de couplage et des historiques de changements.

Ensemble, ces composantes constituent les données de la base de sondage nécessaires pour soutenir les opérations des enquêtes-entreprises. Une entreprise est inscrite dans le RE si elle satisfait à au moins un des critères suivants :

1. elle verse des retenues sur la paye à l'ARC pour ses employés;
2. elle déclare un revenu brut annuel d'au moins 30 000 $;

3. elle est constituée en société en vertu d'une loi fédérale ou provinciale et a produit une Déclaration fédérale de revenus des sociétés (T2) au cours des trois dernières années.

Le RE est établi et tenu à jour à partir de trois principales sources administratives :

- Retenues sur la paye (RP) : versements mensuels ou trimestriels des employeurs à l'ARC pour l'assurance-emploi, les cotisations aux régimes de pensions et autres obligations liées à la paye.
- Comptes de taxe sur les produits et services (TPS) : relevés des versements de TPS et des demandes de crédits de taxe sur les intrants effectués par les entités inscrites.
- Déclarations de revenus : production de déclarations de revenus annuelles des entreprises constituées en société (T2) et non constituées en société (T1), détaillant les revenus, les dépenses et d'autres renseignements financiers.

L'intégration de ces sources administratives est facilitée par le NE comme indiqué à la section 2.1. Le NE sert d'identificateur unique reliant les données sur les RP, la TPS et l'impôt sur le revenu, ce qui réduit les doublons et favorise une représentation cohérente des structures juridiques et opérationnelles au sein du RE.

Le RE fait la distinction entre les *structures juridiques* et les *structures opérationnelles* afin de mieux rendre compte de la complexité du fonctionnement des entreprises. La *structure juridique* définit l'identité juridique de l'entité, qui encadre les interactions avec les institutions gouvernementales (par exemple production de déclarations de revenus et administration de la paye) et façonne les contrats, la propriété et les relations de travail. La *structure opérationnelle* reflète l'organisation interne de l'entreprise, qui comprend la manière d'affecter les ressources, la prestation de biens et services, et le maintien des méthodes comptables.

Pour les petites entreprises, les structures juridiques et opérationnelles sont généralement harmonisées; toutefois, les grandes entreprises fonctionnent souvent par l'entremise de plusieurs entités juridiques couvrant divers secteurs d'activité. Le RE tient compte de cette complexité en reliant plusieurs unités au sein d'une structure consolidée, selon le type de renseignements que chacune peut fournir. Cette approche permet d'adopter des stratégies souples d'échantillonnage et de collecte des données, adaptées à la nature de l'entité et aux objectifs de l'enquête.

2.3 Unités statistiques

Le RE uniformise les structures d'entreprises en *unités statistiques*, bien définies, qui constituent les éléments fondamentaux de la mesure économique et de la conception de l'enquête. Cette transformation est guidée par de grands principes tels que l'autonomie décisionnelle, l'homogénéité de l'activité industrielle et la disponibilité des données. La délimitation de ces unités permet une représentation statistique cohérente de la population des entreprises dans diverses enquêtes et dans divers cadres analytiques.

Le RE repose sur une hiérarchie à quatre niveaux d'unités statistiques. Chaque niveau reflète un aspect différent de la structure organisationnelle et opérationnelle des entreprises et répond à des besoins précis de collecte et d'analyse de données :

- *Entreprise* : il s'agit de l'unité statistique la plus élevée dans la structure opérationnelle et elle représente l'entité juridique complète. L'entreprise correspond à une unité capable de produire un ensemble complet d'états financiers, englobant l'ensemble de ses activités économiques. Elle constitue l'unité principale pour la collecte de renseignements financiers agrégés, y compris les éléments d'actifs, de passif, les revenus et les dépenses.
- *Société* : cette unité correspond au niveau le plus bas auquel le bénéfice d'exploitation peut être calculé. Elle fournit aussi des renseignements détaillés sur les éléments d'actif et de passif, permettant la mesure du capital utilisé. L'unité de la société est particulièrement pertinente pour l'analyse financière et l'évaluation du rendement à un niveau de gestion ou de division.
- *Établissement* : cette unité désigne le plus petit regroupement d'entités de production qui conserve un certain degré d'autonomie organisationnelle et qui exerce ses activités généralement à partir d'un seul emplacement géographique. Les établissements se caractérisent par la production d'un ensemble de biens ou de services relativement homogène. Ils ne s'étendent pas au-delà des frontières provinciales ou territoriales et sont capables de déclarer des données sur les valeurs de sortie, les principaux coûts d'intrants et les dépenses de main-d'œuvre. L'établissement est souvent l'unité privilégiée pour les enquêtes sectorielles et l'analyse de la production.
- *Emplacement* : l'emplacement est une subdivision de l'établissement, représentant un site physique unique ou un groupe défini de sites très près les uns des autres (par exemple une usine, un entrepôt ou un point de vente). Un emplacement peut déclarer certaines variables opérationnelles comme les revenus et, dans de nombreux cas, l'emploi. Ce niveau de détail élevé est particulièrement utile pour les analyses économiques régionales ou locales.

Pour les entreprises complexes, où plusieurs activités ou régions géographiques sont concernées, chaque unité statistique se voit attribuer un code de classification des industries dominant afin d'assurer une catégorisation adéquate. Les entreprises sont classées selon l'industrie primaire associée à leur production globale, tandis que les établissements reçoivent un code de classification des industries et un code de classification géographique reflétant leur activité économique principale, généralement déterminée par les revenus. Ces classifications sont fondées sur le Système de classification des industries de l'Amérique du Nord (SCIAN) et sont régulièrement mises à jour afin de rester pertinentes et exactes.

Le RE accorde la priorité au maintien de l'exactitude de la classification pour les entités commerciales plus grandes et plus complexes, car les erreurs à ces niveaux peuvent avoir des répercussions importantes sur les résultats statistiques. Une surveillance continue et un affinement des classifications des industries et des régions géographiques garantissent l'intégrité et la représentativité de la base de sondage utilisée pour les enquêtes-entreprises.

2.4 Modalités de collecte des données

Les modalités de collecte des données entre l'organisme statistique et une entreprise échantillonnée (définie au niveau de l'unité statistique) sont établies par l'intermédiaire des unités de collecte. Trois attributs sont associés à une unité de collecte :

- *Couverture* : définit la relation entre l'entreprise auprès de laquelle les données sont recueillies et le niveau au sein de l'entreprise (par exemple entreprise, emplacement) pour lequel les données sont requises;
- *Mode de collecte* : le moyen par lequel les données sont obtenues (par exemple questionnaire, interview téléphonique, dossier administratif, etc.);
- *Coordonnées du répondant* : nom, adresse physique ou courriel, et numéro de téléphone au sein de la structure opérationnelle de l'entreprise.

Les figures 2.1 et 2.2 illustrent la façon dont les unités statistiques et les unités de collecte sont respectivement reliées à une entreprise simple ou à une entreprise complexe. Les unités de collecte constituent l'un des moyens de mettre à jour la base de sondage des entreprises en ce qui concerne les données qu'elle contient; d'autres moyens incluent les mises à jour et l'établissement des profils des données administratives. Les unités de collecte représentent aussi le mécanisme permettant de suivre les contacts avec les répondants et d'évaluer le fardeau des répondants. Elles ne sont créées que pour les unités statistiques incluses dans un échantillonnage et sont propres à chaque enquête. Elles sont générées automatiquement selon des règles qui dépendent des liens statistiques opérationnels et de la disponibilité des données. Elles peuvent être modifiées manuellement, au besoin, pour tenir compte de renseignements relatifs à des arrangements particuliers d'établissement de rapports demandés par les répondants.

Figure 2.1 Entreprise simple

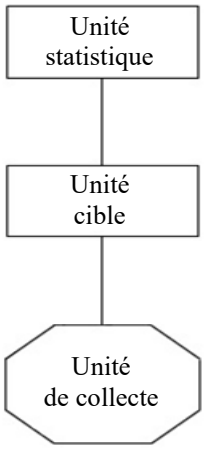
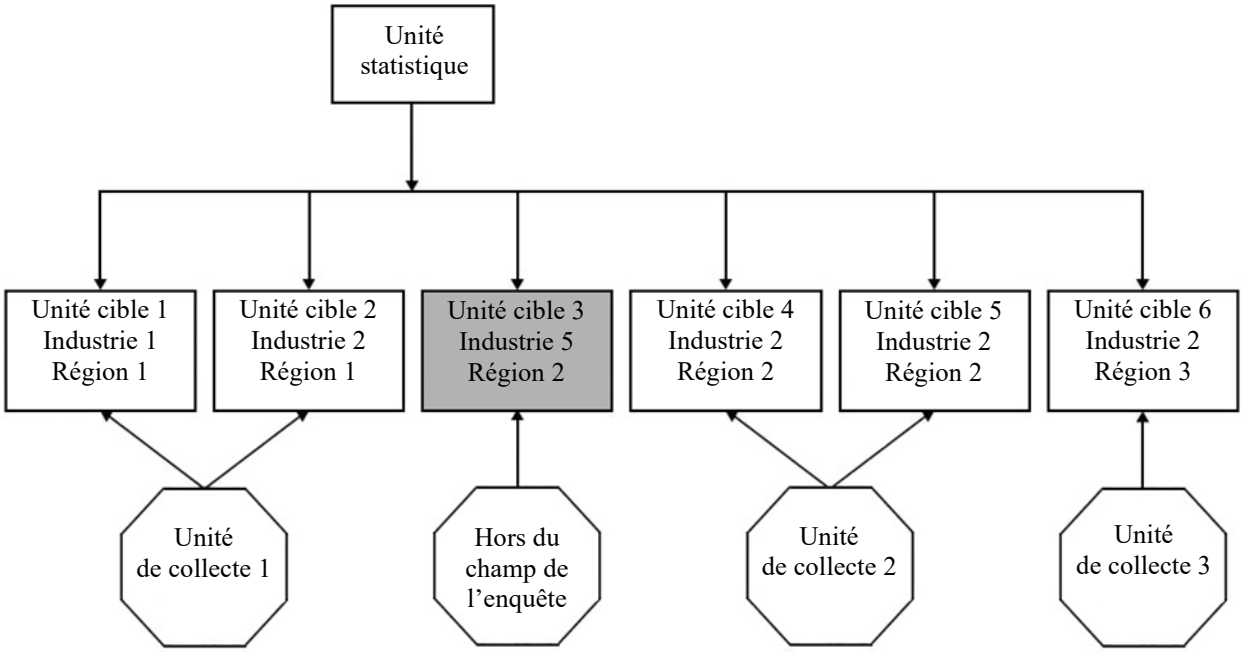


Figure 2.2 Entreprise complexe



2.5 Maintenance

Les entreprises évoluent constamment : ouverture, fermeture, fusion, changement de taille, d'industrie ou d'emplacement. Ces changements peuvent affecter les données de classification, que ce soit dans les parties échantillonnées ou non échantillonnées de la population. Ces modifications peuvent refléter de véritables changements dans les activités de l'entreprise, ou découler d'erreurs de classification antérieures. Les changements d'industrie, d'emplacement ou de taille sont plus souvent décelés chez les entreprises déjà comprises dans un échantillon (faisant partie de l'échantillon) que chez celles qui n'en font pas partie (hors de l'échantillon). Dans les enquêtes uniques, si un changement de classification survient après la sélection de l'échantillon, le poids initial de l'entreprise est conservé aux fins d'estimation, mais la classification mise à jour est utilisée pour l'établissement de rapports. Dans les enquêtes répétées, il peut être nécessaire de reclassifier les entreprises au fil du temps, et l'estimation par domaine peut aider à ajuster ces changements à plusieurs reprises.

Le RE met régulièrement à jour les classifications des industries et des régions géographiques de toutes les entités commerciales, en mettant particulièrement l'accent sur les entreprises de plus grande taille et plus complexes. Le registre conserve également les renseignements sur les entités de production et comptables de ces entreprises.

- *Nouvelles entreprises* : de nouvelles unités entrent dans le RE principalement par l'intermédiaire des comptes de RP, de TPS ou de sociétés. L'ARC attribue un NE à ces unités et demande des renseignements juridiques et opérationnels. Pour les entreprises établies qui acquièrent de nouveaux comptes, ces comptes sont rattachés à leur racine de NE.

- *Disparitions d'entreprises* : les entreprises inactives sont désactivées du RE en fonction de leur activité de versement. Pour les entreprises simples, la désactivation a lieu si elles cessent de faire des versements à l'ARC pendant 12 mois. Pour les entreprises complexes, la désactivation n'intervient que lorsque tous les comptes associés cessent de verser. Si certains comptes cessent de verser, le RE déclenche un nouvel établissement de profils pour mettre à jour la structure et les coordonnées de l'entreprise.
- *Mise à jour de la taille de l'entreprise* : les mesures de la taille de l'entreprise, comme le nombre d'employés et le revenu brut d'entreprise, sont mises à jour à l'aide de modèles selon les versements. Pour les entreprises simples, les données sur les RP et la TPS servent à mettre à jour ces mesures liées à la taille. Pour les entreprises complexes, des enquêtes économiques sont utilisées à cette fin.
- *Enquêtes* : les enquêtes économiques qui reposent sur la base de sondage du RE signalent les changements dans les structures d'entreprises, les coordonnées et les mesures liées à la taille. Le RE examine toute unité présentant des écarts entre les données de la base de sondage et les données recueillies lors de l'enquête. Il examine également les entreprises complexes en réponse à des signaux provenant de sources administratives ou dans le cadre de mises à jour périodiques, comme l'Enquête sur l'établissement de profils. L'établissement de profils est le processus consistant à procéder à des interviews téléphoniques approfondies ou sur place auprès de cadres d'entreprise, en vue d'obtenir toutes les données financières pertinentes, ainsi que les données sur les relations et les structures d'une entreprise (voir College, 1995). Les données recueillies permettent de mettre à jour les entités opérationnelles, juridiques et comptables du RE, ainsi que ses structures statistiques.

3. Calcul de la taille de l'échantillon et répartition

Comme indiqué à la section 2, le RE organise la population des entreprises selon une hiérarchie de quatre unités statistiques : l'entreprise, la société, l'établissement et l'emplacement. Selon les objectifs et les exigences analytiques d'une enquête donnée, l'univers cible peut être défini à l'un ou l'autre de ces niveaux. Une fois l'univers cible délimité, un ensemble correspondant d'unités cibles est déterminé pour former la base du plan de sondage.

L'échantillonnage dans l'univers des entreprises comprend généralement deux étapes principales. La première étape consiste à définir l'univers cible concerné et à déterminer les unités statistiques pertinentes. La deuxième étape consiste à sélectionner l'unité d'échantillonnage, qui doit être au moins aussi détaillée que l'unité cible. Par exemple, si les emplacements et les établissements sont définis comme unités statistiques dans le RE et que l'unité cible est l'emplacement, l'unité d'échantillonnage peut être soit l'emplacement lui-même, soit l'établissement statistique à un niveau plus agrégé. Le choix entre les deux dépend de considérations opérationnelles et méthodologiques. La sélection de l'unité d'échantillonnage a des implications importantes pour la collecte de données, les procédures d'estimation et la logistique globale de l'enquête. Elle influence non seulement le niveau de détail des données recueillies, mais aussi la faisabilité pratique de relier les réponses d'enquête à des sources de donnée administratives ou auxiliaires.

3.1 Stratification des unités d'échantillonnage

Une fois l'unité d'échantillonnage déterminée, la population des entreprises est stratifiée afin d'améliorer l'efficacité et la précision du plan de sondage. Cette stratification est généralement mise en œuvre en deux étapes :

- La stratification primaire repose sur des variables structurelles clés, le plus souvent la *région géographique* (par exemple provinces, territoires et grandes régions métropolitaines) et la *classification des industries* (par exemple commerce de détail, services d'hébergement et de restauration, finance et assurances, ou secteur manufacturier), souvent fondée sur le code SCIAN.
- La stratification secondaire repose sur les mesures de la taille d'une entreprise, pouvant inclure des variables comme le nombre d'employés, le revenu brut annuel ou les ventes totales. Cette stratification tient compte de l'hétérogénéité des unités commerciales ainsi que de la distribution asymétrique vers la droite de l'activité économique, où un petit nombre de grandes entreprises représentent une part disproportionnée de la production économique totale.

Étant donné l'asymétrie marquée des populations des entreprises, la stratification concerne souvent la création d'une strate à tirage complet, dans laquelle toutes les unités sont sélectionnées avec certitude en raison de leur taille ou de leur importance. Les unités restantes sont divisées en une ou plusieurs strates à tirage partiel, à partir desquelles des échantillons sont tirés de façon probabiliste. Deux approches principales sont envisagées :

- Le scénario-n, qui commence par une taille de l'échantillon total fixe de n, est généralement contraint par des limitations budgétaires ou opérationnelles. Dans ce scénario, l'objectif est de répartir l'échantillon disponible de manière optimale entre les strates afin de maximiser la précision dans les limites des ressources allouées.
- Le scénario-c, qui commence par un niveau de précision cible précisé, est exprimé comme un coefficient de variation (c.v.) acceptable pour les estimations clés. Dans ce cas, la taille de l'échantillon requise est calculée pour atteindre le c.v. désiré. Si la taille de l'échantillon obtenue dépasse les ressources disponibles, la cible du c.v. doit être assouplie en conséquence afin de tenir compte des contraintes budgétaires.

Une fois que le scénario-n ou le scénario-c est choisi, les tailles des échantillons pour chaque strate secondaire peuvent être déterminées en utilisant des méthodes de répartition telles que la méthode de répartition de Neyman, la méthode de répartition de puissance ou d'autres techniques optimales conçues pour réduire au minimum la variance ou maximiser l'efficacité, tout en tenant compte des structures de coûts et du fardeau de réponse.

3.1.1 *Scénario-n* : Répartition de la taille globale de l'échantillon aux strates primaires et secondaires

Les tailles d'échantillon sont réparties aux strates primaires au moyen de la méthode de répartition univariée proposée par Bankier (1988). Pour une stratification donnée de l'univers U, supposons que les strates soient désignées par U_{gi}, où $g = 1,\ldots,G$ et $i = 1,\ldots,I$. Supposons que leurs tailles de population respectives soient N_{gi}. L'association de ces strates est représentée par $U = \bigcup_{g=1}^{G} \bigcup_{i=1}^{I} U_{gi}$. Des échantillons aléatoires simples stratifiés s_{gi} de taille n_{gi} sont alors sélectionnés à partir de U_{gi} où $g = 1,\ldots,G$ et $i = 1,\ldots,I$. Les tailles d'échantillon des strates n_{gi} minimisent la fonction de perte :

$$F_1 = \sum_{g=1}^{G} \sum_{i=1}^{I} \left(X_{gi}^{q} \mathrm{CV}(\hat{Y}_{gi}) \right)^2, \tag{3.1}$$

sous la contrainte $n = \sum_{g=1}^{G} \sum_{i=1}^{I} n_{gi}$.

Dans l'équation (3.1), $\mathrm{CV}(\hat{Y}_{gi}) = V(\hat{Y}_{gi}) \big/ Y_{gi}^2$, et X_{gi}^{q} représente une mesure de la taille ou de l'importance de la strate primaire gi^{e} : le paramètre q est une constante comprise dans l'intervalle $0 \le q \le 1$. Le total de la population de la strate primaire gi^{e} est $Y_{gi} = \sum_{\ell \in U_{gi}} y_{gi\ell}$, et $y_{gi\ell}$ est la valeur de la variable ℓ^{e} pour l'unité y dans la strate gi. L'estimateur de Y_{gi} est $\hat{Y}_{gi} = N_{gi} \sum_{\ell \in s_{gi}} y_{gi\ell} \big/ n_{gi}$. On peut démontrer que la fonction F_1 est minimisée par

$$n_{gi} = n \frac{S_{gi} X_{gi}^{q} \big/ \overline{Y}_{gi}}{\sum_{g=1}^{G} \sum_{i=1}^{I} S_{gi} X_{gi}^{q} \big/ \overline{Y}_{gi}}, \tag{3.2}$$

où $V(\hat{Y}_{gi}) = N_{gi}^2 (1/n_{gi} - 1/N_{gi}) S_{gi}^2$, selon $S_{gi}^2 = \sum_{\ell \in U_{gi}} (y_{gi\ell} - \overline{Y}_{gi})^2 \big/ (N_{gi} - 1)$, et $\overline{Y}_{gi} = Y_{gi} / N_{gi}$.

Lorsque $q = 1$, et que l'on suppose que $X_h = Y_h$, on obtient l'allocation de Neyman

$$n_{gi} = n \frac{N_{gi} S_{gi}}{\sum_{g=1}^{G} \sum_{i=1}^{I} N_{gi} S_{gi}}. \tag{3.3}$$

En revanche, lorsque $q = 0$, $S_{gi} \big/ \overline{Y}_{gi}$ ne varient pas de trop d'une strate à l'autre et que les facteurs de corrections peuvent être ignorés, nous obtenons ce qui conduit à une allocation égale entre les strates : c'est-à-dire $n_{gi} = n / GI$. Dans ce cas, les CV des différentes strates sont quasiment égaux. Comme l'a souligné Bankier (1988), l'allocation de Neyman minimise le coefficient de variation (CV) entre les strates primaires. Cependant, certaines strates primaires peuvent avoir des estimateurs des totaux présentant des CV élevés. À l'inverse, une répartition qui permet d'obtenir des CV presque égaux pour les estimateurs au niveau de la strate principale peut donner lieu à des estimateurs globaux ayant un CV beaucoup plus élevé que dans le cadre de la répartition de Neyman.

Bankier (1988) a recommandé $q = 0{,}3$ ou 0,5 comme compromis entre ces deux types d'allocation. D'autres valeurs de q peuvent être utilisées, pourvu qu'elles soient dans l'intervalle [0, 1]. Il convient de mentionner que n_{gi} est fixe pour chaque combinaison de g et de i. Une extension multivariée de la méthode de Bankier (1988) est donnée en annexe A.

En supposant qu'il existe une variable auxiliaire x fortement corrélée à y pour chaque élément de la population, on peut stratifier chaque strate primaire U_{gi}, en strates de taille secondaire. Sans perte de généralité, considérons le cas où nous avons une seule strate primaire. L'extension au cas de plusieurs strates primaires est immédiate, puisque les strates sont indépendantes. À cette fin, supposons qu'une strate primaire donnée soit désignée par U et qu'un échantillon sélectionné à partir de cette strate soit désigné par s. Les tailles de population et d'échantillon correspondantes sont désignées par N et n respectivement. Chaque strate primaire U est alors stratifiée en strate de taille H, $U_1, U_2, \ldots, U_H$ selon les chiffres de population correspondants $N_1, N_2, \ldots, N_H$. Les échantillons sélectionnés dans ces strates sont désignés par s_1, $s_2, \ldots, s_H$, et leurs tailles correspondantes par $n_1, n_2, \ldots, n_H$. Les limites de stratification associées sont désignées par $b_0, b_1, \ldots, b_{H-1}, b_H$, où $b_0 < b_1, \ldots < b_{H-1} < b_H : b_0$ et b_H représentent respectivement les valeurs minimales et maximales de x. Si aucune unité à tirage complet n'est définie, l'estimateur de la moyenne de la population est :

$$\hat{\bar{X}}_1 = N^{-1} \sum_{h=1}^{H} \sum_{k \in s_h} \frac{N_h}{n_h} x_k. \tag{3.4}$$

Pour une strate primaire donnée, les limites de taille sont obtenues en minimisant :

$$V(\hat{\bar{X}}_1) = N^{-2} \left(\sum_{h=1}^{H} \frac{A_h}{n_h} - D \right),$$

où $n_h = n a_h$, $A_h = N_h^2 S_h^2$, $D = \sum_{h=1}^{H} N_h S_h^2$ et S_h^2 est la variance des valeurs connues de x à l'intérieur de la h^{e} strate secondaire contenue dans une strate primaire donnée gi. Le terme a_h est une formule de répartition générale proposée par Hidiroglou et Srinath (1993). Il est donné par

$$a_h = \frac{N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}}{\sum_{h=1}^{H} N_h^{2q_1} \bar{X}_h^{2q_2} S_h^{2q_3}} \quad h = 1, 2, \ldots, H, \tag{3.5}$$

où $0 \le 2q_1 \le 1$, $0 \le 2q_2 \le 1$ et $0 \le 2q_3 \le 1$. Cette formule générale englobe tous les plans de répartition connus, y compris le plan de répartition de puissance proposée par Bankier (1988). Le tableau 3.1 fournit a_h pour les répartitions de Neyman, proportionnelle à X, proportionnelle à N de même que la répartition de puissance proposée par Bankier (1998).

Tableau 3.1
Correspondance entre $a_h, h = 1, 2, \ldots, H$, et plans de répartition connus

Plans de répartition	$a_h, h = 1, 2, \ldots, H,$	$2q_1$	$2q_2$	$2q_3$
Répartition de Neyman	$N_h S_h \Big/ \left(\sum_{h=1}^{H} N_h S_h \right)$	1	0	1
Proportionnelle à X	$X_h \Big/ \left(\sum_{h=1}^{H} X_h \right)$	0	1	0
Proportionnelle à N	$N_h \Big/ \sum_{\ell=1}^{L} N_h$	1	0	0
Répartition de puissance	$X_h^q \Big/ \sum_{h=1}^{H} X_h^q ; 0 \le q \le 1$	q	q	0

En suivant l'approche de Dalenius (1950), l'expression $V(\hat{\bar{X}}_1)$, initialement définie pour une population finie, est transformée en une version applicable à une population continue. Plus précisément, si l'on considère une fonction de densité de probabilité $g(x)$ de la variable auxiliaire x définie sur l'intervalle $(-\infty,\infty)$, la moyenne conditionnelle et la variance de la h^e strate U_h, $h=1,\ldots,H$ peuvent s'écrire $\mu_h = \int_{b_{h-1}}^{b_h} xg(x)\,dx/\omega_h$, et $\sigma_h^2 = \int_{b_{h-1}}^{b_h} x^2 g(x)dx/\omega_h - \mu_h^2$ où $\omega_h = \int_{b_{h-1}}^{b_h} g(x)dx$. La version continue de $V(\hat{\bar{X}}_1)$ s'obtient de la façon suivante :

$$V(\hat{\bar{X}}_1) = \frac{AB}{n} - \frac{D}{N}, \tag{3.6}$$

où $A = \sum_{h=1}^{H} \omega_h^{2q_1}\mu_h^{2q_2}\sigma_h^{2q_3}$, $B = \sum_{h=1}^{H} \left(\omega_h\sigma_h\right)^2 \omega_h^{-2q_1}\mu_h^{-2q_2}\sigma_h^{-2q_3}$ et $D = \sum_{h=1}^{H}\omega_h\sigma_h^2$.

Les limites optimales, $b_1, b_2, \ldots, b_{H-1}, b_H$, où $b_2 \le \ldots \le b_{H-1} \le b_H \le x_{(N)}$, sont déterminées en prenant les dérivées partielles de (3.6) pour chaque $b_h, h=1,\ldots,H$, en en leur attribuant une valeur de zéro, puis en résolvant les équations quadratiques obtenues par une méthode itérative, selon une procédure telle que celle proposée par Sethi (1963). Les valeurs initiales des limites sont fixées en répartissant un nombre égal d'éléments dans chaque groupe. Des précisions supplémentaires sur la procédure de dérivation sont présentées dans Lavallée et Hidiroglou (1988), Baillargeon, Rivest et Ferland (2007), ainsi que Hidiroglou et Kozak (2018). Le programme *R* élaboré par Rivest et Baillargeon (2022) exécute les calculs nécessaires pour obtenir les limites dans ce cas.

Si l'on choisit d'inclure une strate à tirage complet, elle sera désignée par U_H, et les strates restantes à tirage partiel par $U_1, U_2, \ldots, U_{H-1}$. L'estimateur de la moyenne $\bar{X}$ s'écrit alors :

$$\hat{\bar{X}}_2 = N^{-1}\left(\sum_{h=1}^{H-1}\sum_{k\in s_h}\frac{N_h}{n_h}x_k + \sum_{k\in U_H} x_k\right).$$

La variance de population $\hat{\bar{X}}_2$ est donnée par :

$$V(\hat{\bar{X}}_2) = \frac{1}{N^2}\left(\sum_{h=1}^{H-1}\frac{N_h^2}{(n-N_H)a_h}S_h^2 - \sum_{h=1}^{H-1}N_h S_h^2\right). \tag{3.7}$$

L'analogue continu de (3.7) en fonction de la distribution continue $g(x)$ est :

$$V(\hat{\bar{X}}_2) = \frac{AB}{n-N\omega_H} - \frac{D}{N} \tag{3.8}$$

où A, B, et D ont été définis précédemment. Les limites des strates sont obtenues en dérivant l'équation (3.8) par rapport à b_h, $h=1,2,\ldots,H-1$, et en attribuant une valeur de zéro à l'équation ainsi obtenue.

3.1.2 Scénario-*c* : Répartition du coefficient de variation global prévu entre les strates primaires

La taille d'échantillon requise pour chaque strate primaire est déterminée à partir d'un coefficient de variation prévu (c) à l'échelle nationale. Pour une strate primaire donnée $U_{gi}, g=1,\ldots,G,\ i=1,\ldots,I$, supposons que $\hat{X}_{gi} = N_{gi}\sum_{k\in s_{gi}} x_k/n_{gi}$ soit l'estimateur du total de population $X_{gi} = \sum_{k\in U_{gi}} x_k$, où

N_{gi} = nombre d'unités de population,

n_{gi} = nombre d'unités à échantillonner,

et

S_{gi}^2 = variance de la population de x.

Supposons que la variance du total global estimé pour x, $\hat{X}_{..} = \sum_{i=1}^{G}\sum_{i=1}^{I}\hat{X}_{gi}$, soit désignée par $V(\hat{X}_{..})$. Cette variance peut aussi s'écrire : $V(\hat{X}_{..}) = c^2X_{..}^2$, où $X_{..} = \sum_{g=1}^{G}\sum_{i=1}^{I}X_{gi}$ et c est le coefficient de variation global prévu.

Le coefficient de variation national c est réparti entre les strates primaires (province par industrie) au moyen d'une méthode du ratissage (voir Deming et Stephan, 1940). Les étapes sont les suivantes : On calcule les coefficients de variation marginaux pour chaque niveau géographique $c_{g.}$, et chaque niveau industriel $c_{.i}$, en supposant que :

i. Les $c_{g.}$ sont égaux pour $g = 1,\ldots,G$, et

ii. Les $c_{.i}$ sont égaux pour $i = 1,\ldots,I$.

Les $c_{g.}$ et les $c_{.i}$ obéissent alors aux deux conditions suivantes :

iii. $\sum_{g=1}^{G}V(\hat{X}_{g.}) = V(\hat{X}_{..}) = c^2X_{..}^2$,

et

iv. $\sum_{i=1}^{I}V(\hat{X}_{.i}) = V(\hat{X}_{..}) = c^2X_{..}^2$.

où $\hat{X}_{g.} = \sum_{i=1}^{I}\hat{X}_{gi}$, et $\hat{X}_{.i} = \sum_{g=1}^{G}\hat{X}_{gi}$.

On calcule d'abord les coefficients de variation marginaux $c_{g.}$ (coefficient de variation de la g^{e} strate géographique $U_{g.}$) et $c_{.i}$ (coefficient de variation de la i^{e} strate industrielle $U_{.i}$). Puisque $V(\hat{X}_{g.}) = \left(c_{g.}X_{g.}\right)^2$, où $X_{g.} = \sum_{i=1}^{I}X_{gi}$, l'application de la condition (iii) se traduit par $\sum_{g=1}^{G}(c_{g.}X_{g.})^2 = c^2X_{..}^2$. En appliquant la condition (i), on obtient alors :

$$c_{.g} = c\,X_{..}\Big/\sqrt{\sum_{g=1}^{G}\left(X_{.g}\right)^2},\ \text{de}\ g = 1,\ldots,G.$$

De façon analogue, en appliquant les conditions (ii) et (iv), on obtient :

$$c_{.i} = cX_{..}\Big/\sqrt{\sum_{i=1}^{I}X_{.i}^2},\ \text{où}\ X_{.i} = \sum_{g=1}^{G}X_{gi}\ \text{de}\ i = 1,\ldots,I.$$

Un coefficient de variation de la gi^{e} strate primaire est établi à $c_{gi}^{(0)} = 0{,}5\,(c_{g.} + c_{.i})$. La prochaine étape consiste à faire une itération des coefficients de variation comme suit :

$$c_{gi}^{(r)} = c_{gi}^{(r-1)}\begin{cases}\dfrac{\left(c_{.i}X_{.i}\right)}{\sqrt{\sum_{i=1}^{I}c_{gi}^{(r-1)}X_{gi}^2}} & \text{si } r \text{ est impair}\\[2ex] \dfrac{\left(c_{g.}X_{g.}\right)}{\sqrt{\sum_{g=1}^{G}c_{gi}^{(r-1)}X_{gi}^2}} & \text{si } r \text{ est pair}\end{cases}$$

pour $r = 1,\ldots,R,$ jusqu'à ce que le critère de convergence suivant soit satisfait : $\left| c_{gi}^{(t)} - c_{gi}^{(t+1)} \right| < \varepsilon$ pour une certaine $t \geq 5$ et une valeur ε suffisamment petite.

En pratique, cinq itérations suffisent pour stabiliser les valeurs de $c_{gi}^{(r)}$. Les coefficients de variation finaux sont alors désignés $c_{gi}^{(f)}$. On procède ensuite, comme à la section 3.1.1, à la stratification des strates primaires individuelles $gi,$ en conservant la même notation.

Si toutes les strates de taille sont des strates à tirage partiel, l'estimateur de $\bar{X} = N^{-1}\sum_{h=1}^{H}\sum_{k\in U_h} x_k$ est donné par l'équation (3.4). Pour une strate primaire donnée, les limites de taille correspondante sont obtenues en minimisant :

$$n = \frac{N\left(\sum_{h=1}^{H} W_h^2 S_h^2 / a_h\right)}{N(c\bar{X})^2 + \sum_{h=1}^{H} W_h S_h^2}. \tag{3.9}$$

où $W_h = N_h / N$ et $c = c_{gi}^{(f)}$ désigne le coefficient de variation exigé pour la strate primaire $gi,$ et où les autres termes de l'équation (3.9) ont été définis à la section 3.1.1. L'analogue continu de (3.9) s'écrit :

$$n = \frac{N\left(\sum_{h=1}^{H} \omega_h^2 \sigma_h^2 / \alpha_h\right)}{N(c\mu)^2 + \sum_{h=1}^{H} \omega_h \sigma_h^2}$$

où

$$\alpha_h = \frac{\omega_h^{2q_1}\mu_h^{2q_2}\sigma_h^{2q_3}}{\sum_{h=1}^{L-1}\omega_h^{2q_1}\mu_h^{2q_2}\sigma_h^{2q_3}}.$$

Si l'on a une strate à tirage complet et que les autres strates sont à tirage partiel, l'estimateur de $\bar{X} = \sum_{h=1}^{H}\sum_{k\in U_h} x_k / N$ est donné par l'équation (3.6). La taille d'échantillon correspondante est alors :

$$n = N_H + \frac{N\left(\sum_{h=1}^{H-1} W_h^2 S_h^2 / a_h\right)}{N(c\bar{X})^2 + \sum_{h=1}^{H-1} W_h S_h^2}. \tag{3.10}$$

L'analogue continu de (3.10) est :

$$n = N\,\omega_H + \frac{N\,AB}{E}$$

où

$$A = \sum_{h=1}^{H-1} \omega_h^{2q_1}\mu_h^{2q_2}\sigma_h^{2q_3},\ B = \sum_{h=1}^{H-1} (\omega_h \sigma_h)^2\,\omega_h^{-2q_1}\mu_h^{-2q_2}\sigma_h^{-2q_3},\ D = \sum_{h=1}^{H-1} \omega_h \sigma_h^2$$

et $E = N(c\mu)^2 + D.$

Pour une fiabilité donnée c et une répartition $a_h (h = 1,\ldots,H-1),$ supposons que les limites obtenues soient $b_1^{(c)}, b_2^{(c)},\ldots,b_{H-1}^{(c)},$ et que la taille d'échantillon globale minimale associée soit $n.$ En utilisant cette taille d'échantillon n et la même répartition a_h comme données d'entrée pour le scénario $n,$ on obtient les limites $b_1^{(n)}, b_2^{(n)},\ldots,b_{H-1}^{(n)}$ et une fiabilité attendue $c^{(n)}$. Les scénarios c et n sont équivalents, c'est-à-dire $b_h^{(n)} = b_h^{(c)},\ h = 1,\ldots,H-1,$ et $c^{(n)} = c.$ La démonstration figure à l'annexe B.

3.2 Extensions au problème des limites optimales

Les répercussions d'une base de sondage désuète doivent être prises en compte dans la méthode de détermination et de répartition de la taille de l'échantillon. La base de sondage est dite « désuète » lorsque la classification des unités (région géographique, industrie, variable de taille, ou statut « en activité » ou « disparu ») n'est pas à jour. Dans notre cas, on s'intéresse uniquement à l'estimation du total des unités en activité pour une variable y, sachant que les unités disparues figurent dans la base de sondage, mais sont désignées comme actives. Par conséquent, une proportion représentative de ces unités disparues sera incluse dans l'échantillon. L'univers des unités « actives » est classé F_A. L'univers correspondant des unités en activité (mais inconnues) est désigné par F_L $F_L \subset F_{\mathrm{A}}$. Supposons que le paramètre d'intérêt soit le total de domaine $X_L = \sum_{k \in F_A} x_{k\ell}$, où $x_{k\ell}$ est égal à x_k si $k \in F_L$ et à zéro autrement. L'estimateur du total X_L est $\hat{X}_L = \sum_{h=1}^{H} \hat{X}_{hL}$ avec $\hat{X}_{hL} = \sum_{k \in s_{Ah}} \left(N_{Ah} / n_{Ah} \right) x_{k\ell}$. Il s'agit de déterminer la taille d'échantillon n_A, de telle sorte que : i) la répartition de l'échantillon entre les strates du plan de sondage soit $n_{Ah} = n_A a_h (0 < a_h < 1)$; ii) le coefficient de variation visé c soit satisfait, c'est-à-dire que $V(\hat{X}_L) = c^2 X_L^2$. La population F_A est stratifiée en H strates $F_{A,h}$, chacune de taille N_{Ah}. Des échantillons aléatoires simples s_{Ah} de taille n_{Ah} sont sélectionnés dans chaque strate $F_{A,h}, h = 1, \ldots, H$, sans remplacement. Supposons que $F_{L,h}, h = 1, \ldots, H$ désigne l'ensemble correspondant des unités en activité dans la strate $F_{A,h}$. Définissons

$$V(\hat{X}_L) = c^2 X_L^2 = \sum_{h=1}^{H} N_{Ah}^2 \left(\frac{1}{n_{Ah}} - \frac{1}{N_{Ah}} \right) S_{Lh}^2$$

où $S_{Lh}^2 = (N_{Ah} - 1)^{-1} \sum_{k \in F_{A,h}} (x_{k\ell} - \bar{X}_{hL})^2$ et $\bar{X}_{hL} = \sum_{k \in F_{A,h}} x_{k\ell} / N_{Ah}$.

En utilisant $V(\hat{X}_L)$, la taille d'échantillon globale requise n_A est

$$n_A = \frac{\sum_{h=1}^{H} N_{Ah}^2 \left(\tilde{S}_{Lh}^2 + (1 - P_{Lh}) \tilde{\bar{X}}_{Lh}^2 \right) P_{Lh} / a_h}{c^2 \left(\sum_{h=1}^{H} N_{Ah} \tilde{\bar{X}}_{Lh} P_{Lh} \right)^2 + \sum_{h=1}^{H} N_{Ah} \left(\tilde{S}_{Lh}^2 + (1 - P_{Lh}) \tilde{\bar{X}}_{Lh}^2 \right) P_{Lh}}$$

où $\tilde{S}_{Lh}^2 = (N_{Lh} - 1)^{-1} \sum_{k \in F_{hL}} (x_k - \tilde{\bar{Y}}_{Lh})^2$, $\tilde{\bar{Y}}_{Lh} = \sum_{k \in F_{Ah}} x_{k\ell} / N_{Lh}$, N_{Lh} est le nombre d'unités appartenant au domaine $F_{Lh} = F_L \cap F_{Ah}$, et $P_{Lh} = N_{Lh} / N_{Ah}$ est la proportion attendue d'unités appartenant à U_ℓ et qui ont été échantillonnées au départ dans la strate h.

Il convient de mentionner que lorsque $P_{h\ell} = 1$, nous obtenons l'équation (3.9). Les composantes de moyenne et de variance peuvent être estimées à partir d'enquêtes antérieures. Les tailles d'échantillon requises au niveau de la strate sont alors simplement $n_{Ah} = n_A a_h$ pour $h = 1, \ldots, H$. Il n'est pas recommandé d'utiliser une approximation du type $n_{Ah}^* = n_{Ah} a_h$ pour compenser la présence d'unités disparues inconnues lorsque n_h est calculée en ignorant leur existence dans l'univers.

Rivest (2002) a élargi l'algorithme de stratification de Lavallée et Hidiroglou (1988) en faisant la distinction entre la variable de stratification et la variable d'enquête. Il a présenté un algorithme de stratification selon un modèle log-linéaire, qu'il a appliqué à plusieurs enquêtes élaborées par le service de consultation statistique de l'Université Laval. Baillargeon et Rivest (2009) ont approfondi ce travail en intégrant

la possibilité d'une strate à tirage nul et la prise en compte de probabilités uniformes de non-réponse entre les strates de taille. De plus, ils ont démontré que, d'un point de vue computationnel, l'algorithme proposé par Kozak (2004) constituait l'outil le plus efficace pour résoudre le problème d'optimisation lié à la détermination des limites de strates. D'autres travaux de Hidiroglou et Laniel (2001) ainsi que de Lednicki et Wieczorkowski (2003) ont proposé d'intégrer dans le processus de stratification les moments anticipés de la variable d'intérêt y, conditionnellement à une variable auxiliaire x. Il devrait y avoir un nombre minimal d'unités dans chaque strate à tirage partiel, afin de garantir que la non-réponse ne conduise pas à des strates vides. Un poids maximal doit être déterminé afin d'éviter l'incidence négative des valeurs aberrantes. L'application des règles supplémentaires ci-dessus est clairement illustrée par Ferland et Batten (2007).

3.3 Comparaison numérique entre les répartitions dans les strates primaires

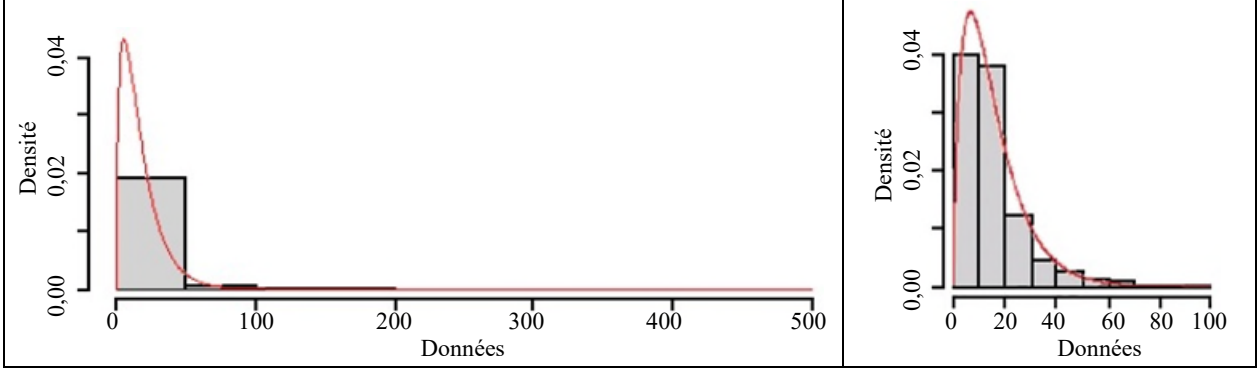
Les méthodes présentées dans les sections 3.1.1 et 3.1.2 sont illustrées en créant des ensembles de données selon celui donné dans Rivest et Baillargeon (2022). Cet ensemble de données a été créé à partir d'une mesure de la taille simulée de l'Enquête mensuelle sur le commerce de détail (EMCD) menée par Statistique Canada. Cet ensemble de données comportait 2 000 établissements. La mesure de la taille de chaque établissement a été obtenue en combinant les données de l'EMCD avec trois variables administratives issues des déclarations de revenus des sociétés. En raison de la nature hautement asymétrique des données, les 21 plus grandes observations ont été supprimées afin d'obtenir un ajustement convergent de la distribution gamma. La distribution gamma est donnée par $f(x) = (\Gamma(\alpha))^{-1}\beta^{\alpha}x^{\alpha-1}e^{-\beta x}$, où α et β sont respectivement les paramètres de forme et de taux. Ces paramètres ont été estimés comme suit : $\hat{\alpha} =$ 1,7610 (0,0515) et $\hat{\beta} =$ 0,1147 (0,0039), où les valeurs entre parenthèses correspondent aux erreurs types associées aux estimateurs. La figure 3.1 présente les statistiques sommaires pour les données originales ainsi que pour les données tronquées.

Pour évaluer l'incidence des différentes méthodes de répartition sur le plan de sondage stratifié, une comparaison numérique a été réalisée sur quinze strates primaires. Ces strates ont été définies par une classification croisée de cinq provinces canadiennes et de trois groupes d'industries. Le choix de ces provinces et de ces industries repose sur leur importance économique et leur diversité sectorielle, ce qui permet des comparaisons significatives entre strates. Les provinces comprenaient le Québec, l'Ontario, le Manitoba, l'Alberta et la Colombie-Britannique. Les groupes d'industries étaient fondés sur le niveau à trois chiffres du SCIAN. Les codes du SCIAN sélectionnés étaient :

- 441 : Concessionnaires de véhicules et de pièces automobiles,
- 444 : Marchands de matériaux de construction et de matériel et fournitures de jardinage,
- 445 : Magasins d'alimentation.

Figure 3.1 Statistiques sommaires

Données originales de l'EMCD	Données tronquées de l'EMCD
n = 2 000 : coefficient d'asymétrie = 8,622	n = 1 979 : coefficient d'asymétrie = 2,315
Densités empiriques et théoriques	Densités empiriques et théoriques



Note: EMCD = Enquête mensuelle sur le commerce de détail.

Les tableaux 3.2 et 3.3 présentent respectivement le nombre d'établissements, ainsi que les ventes totales pour les quinze strates primaires.

Tableau 3.2
Nombre d'établissements de commerce de détail (N_{gi}) pour chaque strate primaire, $g = 1,\ldots,5$ et $i = 1, 2, 3$

Province	Industrie			
	SCIAN 441	SCIAN 444	SCIAN 445	Ensemble des industries
Québec	3 301	1 826	7 108	12 235
Ontario	4 602	2 821	9 980	17 403
Manitoba	511	336	839	1 686
Alberta	555	834	3 707	5 096
Colombie-Britannique	1 826	1 220	3 707	6 753
Ensemble des provinces	10 795	7 037	25 341	43 173

Note: SCIAN = Système de classification des industries de l'Amérique du Nord.
Source : Extrait de Statistique Canada. Tableau 33-10-0761-01 Nombre d'entreprises canadiennes, avec employés, juin 2024 https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=3310076101.

Tableau 3.3
Ventes totales (X_{gi}) en milliers de dollars (non désaisonnalisées) pour chaque strate primaire $g = 1,\ldots,5$ et $i = 1, 2, 3$

Province	Industrie			
	SCIAN 441	SCIAN 444	SCIAN 445	Ensemble des industries
Québec	4 581 326	1 130 512	3 211 742	8 923 580
Ontario	7 107 009	1 568 757	4 850 610	13 526 376
Manitoba	642 794	169 960	471 349	1 284 103
Alberta	2 505 955	570 628	1 509 526	4 586 109
Colombie-Britannique	1 969 426	567 892	1 992 689	4 530 007
Ensemble des provinces	16 806 510	4 007 749	12 035 916	32 850 175

Note: SCIAN = Système de classification des industries de l'Amérique du Nord.
Source : Extrait de Statistique Canada. Tableau 20-10-0056-01 Ventes mensuelles du commerce de détail par province et territoire (× 1 000), juin 2024.

Ensuite, les coefficients de variation pour les totaux provinciaux par industrie ont été calculés. Ces coefficients ont été obtenus à l'aide de la méthode du ratissage décrite à la section 3.1.2. Le coefficient de

variation global c a été fixé à 1 %. Les coefficients de variation qui en découlent sont présentés dans le tableau 3.4.

Tableau 3.4
Coefficients de variation redressés par la méthode itérative du quotient $c_{gi}^{(f)}$ (%), selon la province et l'industrie, $g = 1,\ldots,5$ et $i = 1, 2, 3$

Province	Industrie			
	SCIAN 441	SCIAN 444	SCIAN 445	Ensemble des industries
Québec	2,89 %	2,92 %	2,92 %	1,89 %
Ontario	2,88 %	2,91 %	2,91 %	1,89 %
Manitoba	2,92 %	2,95 %	2,95 %	1,89 %
Alberta	2,86 %	2,89 %	2,89 %	1,89 %
Colombie-Britannique	2,91 %	2,94 %	2,94 %	1,89 %
Ensemble des provinces	1,54 %	1,54 %	1,54 %	1,00 %

Note: SCIAN = Système de classification des industries de l'Amérique du Nord.

Les observations individuelles pour chacune de ces strates primaires ont été générées à partir des données d'entrée suivantes :

1. le nombre d'établissements (comme indiqué au tableau 3.2);
2. leurs ventes totales (comme indiqué au tableau 3.3);
3. l'estimation des paramètres de la distribution gamma où $\hat{\alpha} = 1{,}7610$ et $\hat{\beta} = 0{,}1147$.

Les valeurs générées ont ensuite été réparties proportionnellement pour s'assurer que leurs agrégats correspondent aux chiffres des ventes totales rapportés au tableau 3.3. En utilisant ces valeurs générées, les strates primaires ont été stratifiées selon trois méthodes de répartition de l'échantillon. Il s'agissait de :

1. deux strates à tirage partiel, au moyen de la répartition de Neyman;
2. une strate précise à tirage complet ainsi qu'une seule strate à tirage partiel;
3. une strate précise à tirage complet combinée à deux strates à tirage partiel, au moyen de la répartition de Neyman.

La troisième méthode a été retenue, car elle aboutissait au plus petit nombre d'unités dans tous les cas.

En utilisant les coefficients de variation du tableau 3.4 pour chaque strate primaire, les tailles d'échantillon requises sont présentées au tableau 3.5.

Tableau 3.5
Tailles d'échantillon (n_{gi}) selon la province et l'industrie avec strate précise à tirage complet et deux strates à tirage partiel (répartition de Neyman). $g = 1,\ldots,5$ et $i = 1, 2, 3$

Province	Industrie			
	SCIAN 441	SCIAN 444	SCIAN 445	Ensemble des industries
Québec	183	170	194	547
Ontario	189	175	201	565
Manitoba	139	90	134	363
Alberta	167	130	191	488
Colombie-Britannique	157	147	185	489
Ensemble des provinces	835	712	905	2 452

Note: SCIAN = Système de classification des industries de l'Amérique du Nord.

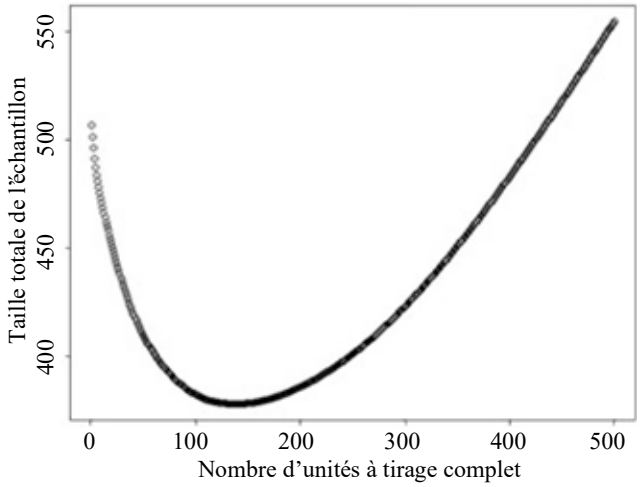
Il convient de mentionner que la spécification d'une strate à tirage complet combinée à une seule strate à tirage partiel peut entraîner une taille d'échantillon globale plus élevée que l'utilisation de deux strates à

tirage partiel combinée à la répartition de Neyman. Par exemple, la taille d'échantillon pour les concessionnaires de véhicules et de pièces automobiles en Alberta aurait été de 191 selon deux strates à tirage partiel utilisant la répartition de Neyman, alors qu'elle aurait été de 378 unités selon une combinaison d'une strate à tirage complet et d'une strate à tirage partiel. Ce résultat, illustré à la figure 3.2, a été confirmé à l'aide de la formule exacte fournie par Hidiroglou (1986b), donnée par :

$$n(t) \;=\; N - \frac{(N-t)\,c^2X^2}{c^2X^2 + (N-t)\,S^2_{[N-t]}}$$

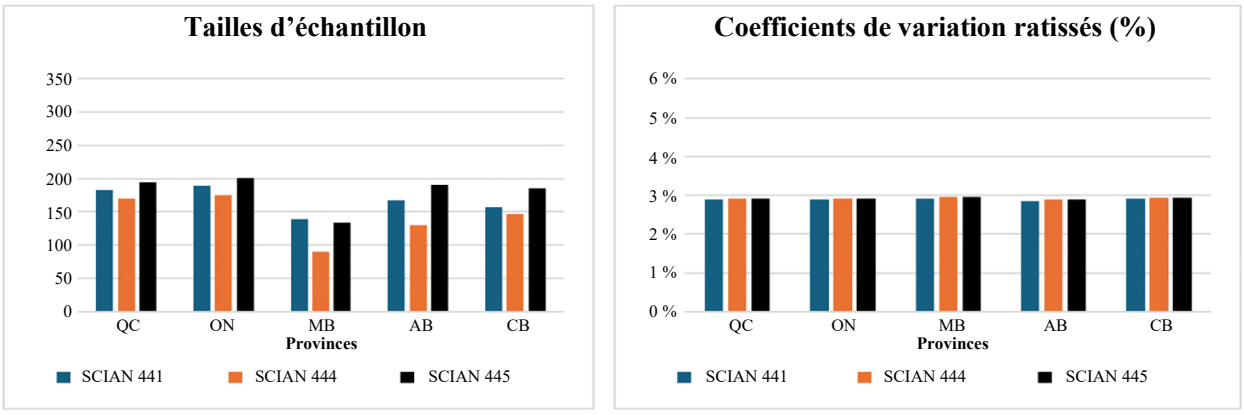
où $n(t)$ représente le nombre total d'unités à échantillonner (y compris les unités à tirage complet), t est le nombre d'unités à tirage complet; $S^2_{[N-t]}$ est la variance de la population des plus petites unités $N-t$, et c est le coefficient de variation précisé.

Figure 3.2 Taille de l'échantillon par rapport au nombre d'unités à tirage complet



Une représentation graphique des résultats des tableaux 3.4 et 3.5 est présenté dans la figure 3.3.

Figure 3.3 Strates primaires : tailles d'échantillon (n_{gi}), et coefficients de variation ratissés correspondants ($c_{gi}^{(f)}$), $g = 1,\ldots,5$ et $i = 1,2,3$



Note: SCIAN = Système de classification des industries de l'Amérique du Nord.

Ensuite, les tailles d'échantillon $n_{gi} = nX_{gi}^q \Big/ \sum_{g=1}^{G}\sum_{i=1}^{I} X_{gi}^q$, $g = 1,\ldots,5$ et $i = 1, 2, 3$ où $0 \le q \le 1,$ ont été calculées à l'aide de la répartition de puissance de Bankier selon l'équation (3.3). La taille d'échantillon totale pour les 15 strates primaires était de $n = 2\,452$ (voir tableau 3.5). Cela permet de comparer les coefficients de variation et les fractions d'échantillonnage obtenus pour chacune des 15 strates primaires avec ceux obtenus à l'aide de la méthode du ratissage résumée à la figure 3.3.

Les deux valeurs extrêmes pour q sont 0 et 1. Lorsque $q = 0,$ nous avons une répartition égale de l'échantillon à chacune des strates primaires : c'est-à-dire $n_{gi} = n / N \quad g = 1,\ldots,G$ et $i = 1,\ldots,I.$ Lorsque $q = 1,$ nous avons une répartition de Neyman de l'échantillon à chacune des strates primaires : c'est-à-dire $n_{gi} = n / N \quad g = 1,\ldots,G$ et $i = 1,\ldots,I$: c'est-à-dire

$$n_{gi} = n\left(\sum_{g=1}^{G}\sum_{i=1}^{I} N_{gi}S_{gi}\right)^{-1} N_{gi}S_{gi} \quad g = 1,\ldots,G \text{ et } i = 1,\ldots,I.$$

Comme nous le verrons plus loin, aucun de ces choix n'est idéal. La stratification des 15 strates primaires en strates secondaires a utilisé la même stratification par taille que celle présentée dans la table 3.5 : c'est-à-dire une strate à tirage complet et deux strates à tirage partiel (en utilisant la répartition de Neyman).

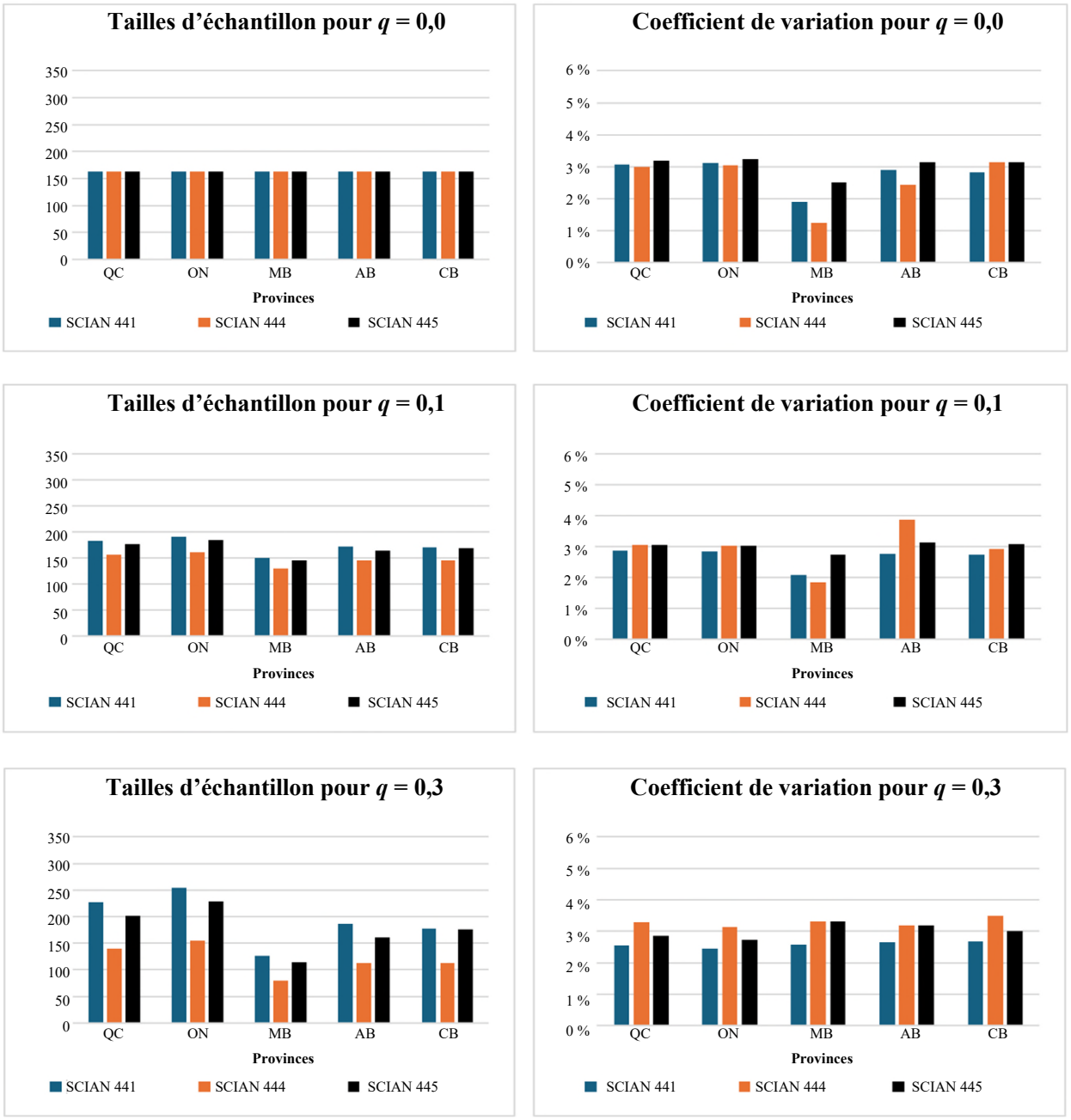
Nous avons comparé les fractions d'échantillonnage et les coefficients de variation à l'aide de la méthode du ratissage avec cinq valeurs différentes pour la sélection des valeurs de q (0,0; 0,1; 0,3; 0,5; et 1,0) dans la procédure de Bankier (1988). Ces comparaisons reposaient sur une taille d'échantillon totale de 2 452, calculée à l'aide de la méthode du ratissage décrite à la section 3.2.2. Un résumé graphique des résultats est présenté à la figure 3.4 pour les différentes valeurs de q.

Plusieurs observations peuvent être tirées de la figure 3.4 :

- *Tailles d'échantillon* : Les tailles d'échantillon augmentent à mesure que q augmente dans les provinces où les ventes sont les plus élevées, le Québec et l'Ontario. L'augmentation de la taille d'échantillon est modérée dans les deux provinces suivantes où les ventes sont les plus élevées à mesure que q augmente, l'Alberta et la Colombie-Britannique. Cela est vrai pour les codes SCIAN 441 et 445. Cependant, elles diminuent pour le code SCIAN 444 à mesure que q augmente dans toutes les provinces. La taille de l'échantillon diminue à mesure que q augmente pour la plus petite province, le Manitoba, pour tous les codes SCIAN. Cela est vrai pour la méthode du ratissage (figure 3.3) et la répartition de Bankier (1988) (figure 3.4). Lorsque $q = 0,$ la taille de l'échantillon est la même pour toutes les strates primaires, soit 163. Lorsque $q = 1,$ la taille de l'échantillon augmente à mesure que la taille des strates primaires augmente.
- *Coefficients de variation* : Les coefficients de variation diminuent à mesure que q augmente dans les provinces où les ventes sont les plus élevées, le Québec et l'Ontario. Les coefficients de variation diminuent modérément dans les deux provinces suivantes, l'Alberta et la Colombie-Britannique, où les ventes sont les plus élevées, à mesure que q augmente. Cela est vrai pour les codes SCIAN 441 et 445. Cependant, ils augmentent pour le code SCIAN 444 à mesure que q augmente dans toutes les provinces. Les coefficients de variation diminuent à mesure que q augmente pour la plus petite province, le Manitoba, pour tous les codes SCIAN. Cela est vrai pour la méthode du ratissage (figure 3.3) et la répartition de Bankier (1988) (figure 3.4). Lorsque $q = 0,$ les coefficients de variation varient entre 1,25 % et 3,12 %. Lorsque $q = 1,$ les coefficients de variation diminuent à mesure que la taille des strates primaires augmente.

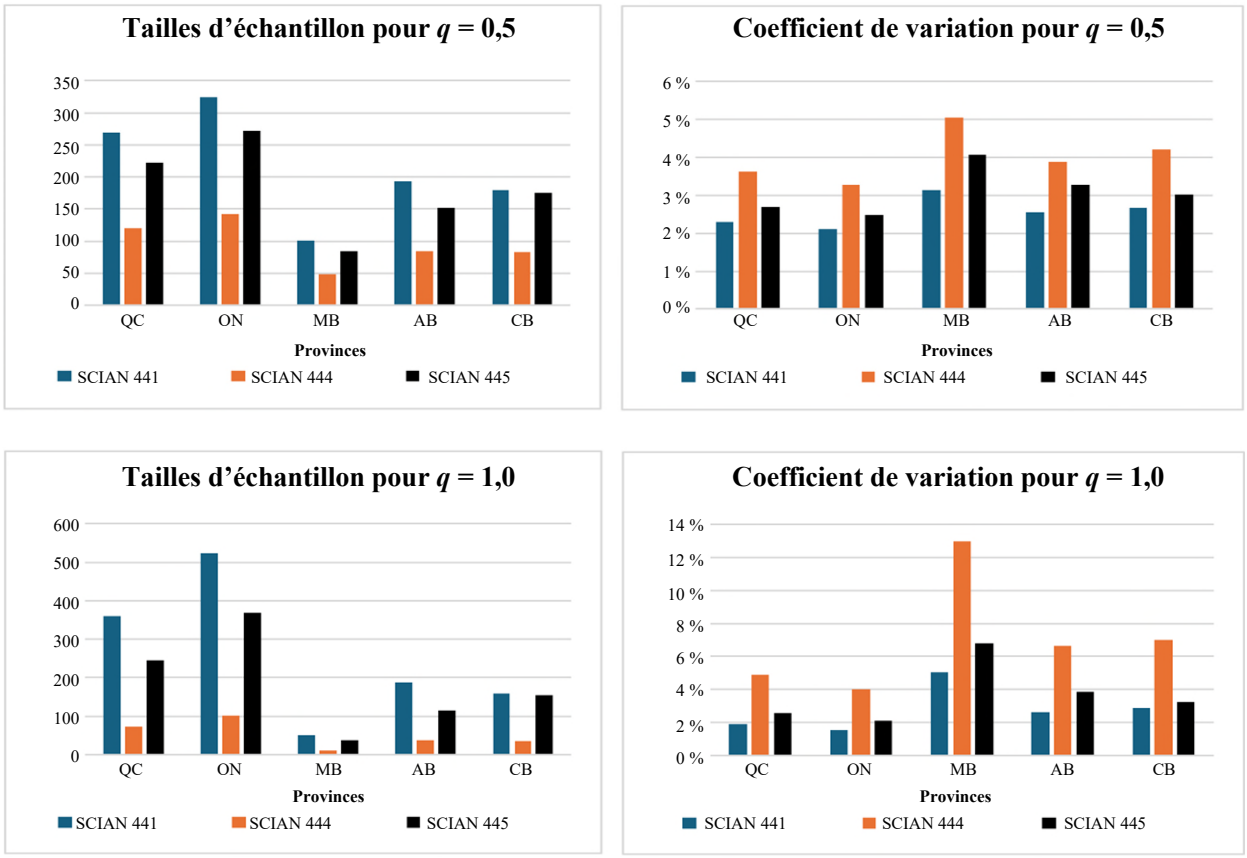
Les valeurs de q comprises entre 0,1 et 0,3 offrent un compromis raisonnable entre la taille des échantillons et les CV. Bien que la méthode de ratissage produise des CV comparables entre les quinze strates primaires, elle nécessite des tailles d'échantillon plus importantes dans les strates où les ventes sont plus faibles. Les résultats du scénario c (figure 3.3) sont proches de ceux obtenus par le scénario n (figure 3.4), lorsque $q = 0,1$.

Figure 3.4 Tailles d'échantillon des strates primaires (n_{gi}) obtenues à l'aide de diverses valeurs q (0,0; 0,1; 0,3; 0,5 et 1,0) et coefficients de variation correspondants (c_{gi}), $g = 1,\ldots,5$ et $i = 1, 2, 3$



Note: SCIAN = Système de classification des industries de l'Amérique du Nord.

Figure 3.4 (suite) Tailles d'échantillon des strates primaires (n_{gi}) obtenues à l'aide de diverses valeurs q (0,0; 0,1; 0,3; 0,5 et 1,0) et coefficients de variation correspondants (c_{gi}), $g = 1,\ldots,5$ et $i = 1, 2, 3$



Note: SCIAN = Système de classification des industries de l'Amérique du Nord.

4. Renouvellement et coordination des échantillons

4.1 Introduction

Dans la présente section, nous passons en revue les méthodes de sélection initiale de l'échantillon, de renouvellement de l'échantillon et de sélection des nouvelles entreprises (créations) dans les enquêtes-entreprises. Ces méthodes visent à garantir que l'échantillon reflète toujours la structure de la population au fil du temps, permettant ainsi de produire des estimations précises et fiables. Les objectifs principaux de ces méthodes sont de réduire le fardeau de réponse pour les entreprises, en particulier les plus petites, et d'améliorer la qualité des estimations, telles que les variations et les moyennes annuelles.

Le renouvellement de l'échantillon consiste à remplacer périodiquement une partie de l'échantillon tout en en conservant une partie. Cette stratégie contribue à réduire le fardeau des entreprises. Un aspect essentiel du renouvellement de l'échantillon est le respect des exigences en matière « d'inclusion » et « d'exclusion ». Ces contraintes sont conçues pour s'assurer que les entreprises ne sont pas sursollicitées par l'enquête, tout en maintenant l'intégrité de l'échantillonnage. L'exigence en matière d'inclusion correspond à la durée

minimale pendant laquelle une entreprise doit rester dans l'échantillon. Par exemple, une période typique de temps d'inclusion pourrait être de 12 mois. L'exigence en matière d'exclusion stipule que les entreprises retirées de l'échantillon ne doivent pas être réintégrées avant une période déterminée, généralement 12 mois. Cela permet d'éviter la surenquête, qui peut engendrer une fatigue en matière de réponse et éventuellement fausser les données.

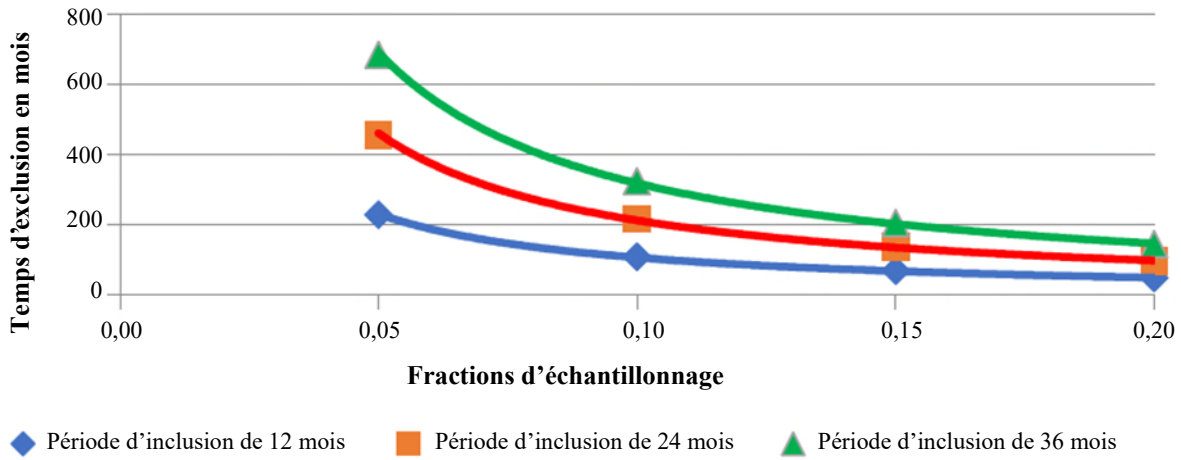
Dans la pratique, certaines unités peuvent devoir rester plus longtemps dans l'échantillon que la durée minimale prescrite pour la période d'inclusion afin de satisfaire les deux exigences et de préserver l'intégrité du plan de sondage. Bien que cela puisse légèrement accroître le fardeau de réponse pour ces entreprises, cette solution reste généralement plus efficace que de les réintégrer trop rapidement après les avoir exclues.

Pour une strate à tirage partiel donnée U_h, supposons que le taux d'échantillonnage soit f_h. Définissons $t_{h,\text{in}}$ et $t_{h,\text{out}}$ comme le nombre de fois où une unité est respectivement incluse et exclue de l'échantillon. La relation suivante s'applique $t_{h,\text{out}} = f_h^{-1}(1-f_h)\,t_{h,\text{in}}$. La figure 4.1 présente un graphique où l'axe des X correspond aux fractions d'échantillonnage, et l'axe des Y correspond au temps d'exclusion. Le graphique inclut plusieurs courbes (ou points de données) correspondant à différentes valeurs de temps d'inclusion.

À partir de la figure 4.1, nous observons que :

1. Il existe une relation positive entre le temps d'inclusion et le temps d'exclusion. À mesure que le temps alloué au processus actif (inclusion) augmente, le temps passé dans la phase inactive (exclusion) augmente également.
2. Pour toute valeur fixe du temps d'inclusion, la période d'exclusion tend à diminuer de manière exponentielle à mesure que la fraction d'échantillonnage augmente. Cela laisse entendre que des fractions d'échantillonnage plus élevées entraînent des transitions plus rapides ou plus efficaces entre les périodes actives et inactives.
3. Pour une fraction d'échantillonnage constante, alors que le temps d'inclusion augmente, le temps d'exclusion augmente aussi. Cela indique que lorsque plus de temps est passé dans le processus actif, un temps d'exclusion plus long est alors requis, quelle que soit la fraction d'échantillonnage.

Figure 4.1 Relation entre le temps d'inclusion et le temps d'exclusion pour des taux d'échantillonnage donnés



Il existe deux principaux types de scénarios de renouvellement utilisés dans les enquêtes-entreprises : *le scénario de renouvellement dépendant du plan d'enquête* et *le scénario de renouvellement non dépendant du plan d'enquête*.

1. Le scénario de renouvellement *dépendant du plan d'enquête* ne permet de renouveler les unités échantillonnées qu'à l'intérieur des strates d'une seule enquête. Cependant, il ne gère pas et ne contrôle pas le chevauchement des unités échantillonnées entre différentes enquêtes.
2. Le scénario de renouvellement *indépendant du plan d'enquête*, d'autre part, permet à la fois le renouvellement des unités à l'intérieur d'une enquête donnée et le contrôle du chevauchement des échantillons entre plusieurs enquêtes. Ce chevauchement contrôlé est appelé *coordination d'échantillons*. Celle-ci est essentielle pour réduire le fardeau de réponse et améliorer l'efficacité lorsque plusieurs enquêtes sont tirées d'une même base de sondage.

Il existe plusieurs raisons de mettre en œuvre la coordination d'échantillons dans les enquêtes-entreprises. Dans certains cas, l'objectif est de *réduire le fardeau de réponse*, qui se concrétise par la *coordination négative*, c'est-à-dire en établissant la sélection de l'échantillon de manière à éviter d'inclure la même entreprise dans plusieurs enquêtes sur une courte période. Dans d'autres situations, l'objectif peut au contraire viser à *maximiser le chevauchement des échantillons* entre les enquêtes, connu sous le nom de *coordination positive*. Cette approche favorise la continuité lors du remaniement des enquêtes, réduit les coûts de transition et facilite l'utilisation des données d'une enquête pour en soutenir une autre. Une *approche hybride*, combinant à la fois la coordination positive et la coordination négative, est aussi possible. Par exemple, on peut chercher à maintenir un fort chevauchement des échantillons successifs d'une enquête mensuelle donnée (coordination positive), tout en limitant la durée pendant laquelle une unité demeure dans l'échantillon afin d'éviter d'imposer un fardeau trop élevé aux répondants (coordination négative).

Nous présentons ensuite deux exemples de scénarios de renouvellement dépendants du plan d'enquête, afin d'illustrer de quelle façon ces principes s'appliquent en pratique.

4.2 Scénarios de renouvellement dépendant du plan d'enquête

Du début au milieu des années 1980, Statistique Canada a mis en place des scénarios de renouvellement pour deux de ses enquêtes-entreprises phares : l'Enquête sur l'emploi, la rémunération et les heures de travail (EERH) et l'Enquête mensuelle sur le commerce de gros et de détail (EMCGD). L'EERH fournit un portrait statistique mensuel de la rémunération des emplois, du nombre d'emplois (c'est-à-dire postes occupés) et des heures travaillées selon des catégories d'industrie détaillées à l'échelle du pays, des provinces et des territoires. L'EMCGD, quant à elle, recueille des données mensuelles sur les ventes au détail et en gros ainsi que sur les stocks, ventilées par industrie et par région géographique. Au moment de leur mise en œuvre, les deux enquêtes s'appuyaient sur la Classification type des industries, 1980 (CTI), qui fournissait un cadre cohérent pour catégoriser les activités économiques dans l'ensemble du Canada.

L'EERH a mis en place une méthode de renouvellement connue sous le nom de groupe de renouvellement échantillonné, tandis que l'EMCGD a adopté une approche différente appelée échantillonnage par panel. Malgré l'utilisation de techniques distinctes, les deux enquêtes ont appliqué des stratégies de renouvellement alignées sur le plan stratifié de leurs échantillons. Plus précisément, le renouvellement des unités échantillonnées était effectué à l'intérieur de chaque strate secondaire de taille, de manière à préserver l'intégrité de la stratification. Ces deux enquêtes sont encore opérationnelles aujourd'hui. Des renseignements récents sur l'EERH sont disponibles dans EERH (2025). À l'heure actuelle, l'EMCGD est menée comme deux enquêtes distinctes : l'Enquête mensuelle sur le commerce de détail (EMCD) (voir EMCD, 2025) et l'Enquête mensuelle sur le commerce de gros (EMCG) (voir EMCG, 2025).

Voici un résumé des méthodes utilisées dans chaque scénario de renouvellement pour l'EERH et l'EMCGD au début des années 1990.

4.2.1 Groupes de renouvellement

Le scénario de renouvellement de l'EERH a été élaboré au début des années 1980. Un aperçu de ce scénario est présenté ci-dessous. L'échantillon mensuel était constitué de 14 groupes, numérotés de 0 à 13. Le groupe 0 comprenait les unités à tirage complet dans une strate primaire. Les groupes 1 à 13 étaient désignés comme des groupes de renouvellement. Les numéros 1 à 12 des groupes de renouvellement correspondaient au mois d'entrée des unités (autres que les créations) dans l'échantillon. Par exemple, le groupe de renouvellement 1 comprenait principalement les unités entrées dans l'échantillon en janvier, ainsi que toutes les nouvelles unités (créations), alors que le groupe 2 contenait les unités entrées en février, et ainsi de suite. Le groupe de renouvellement 13 était composé des unités présentes dans l'échantillon depuis 12 mois : il s'agissait des unités les plus anciennes en matière de temps d'inclusion dans l'échantillon. Ces unités devenaient admissibles pour être retirées de l'échantillon. Chaque mois, de nouvelles créations étaient sélectionnées aléatoirement et réparties dans les groupes de renouvellement. Des détails sur le scénario de renouvellement utilisé pour l'EERH sont fournis dans Şchiopu-Kratina et Srinath (1991).

4.2.2 Échantillonnage par panel

L'EMCGD reposait sur un scénario de renouvellement fondé sur les panels. Dans ce scénario, la population de chaque strate secondaire de taille était divisée aléatoirement en groupes de renouvellement, en veillant à ce que le nombre d'unités dans chaque groupe diffère d'au plus d'une unité. Le nombre total de groupes de renouvellement était déterminé en fonction des fractions d'échantillonnage et des contraintes d'inclusion et d'exclusion. Un échantillon aléatoire simple de groupes de renouvellement était sélectionné, avec une proportion de groupes choisie de manière à correspondre à la fraction d'échantillonnage désirée pour chaque strate. L'échantillon comprenait alors toutes les unités provenant des groupes de renouvellement sélectionnés. Le renouvellement avait lieu en ajoutant de nouvelles unités (créations) aux groupes de façon systématique, tout en retirant certains groupes faisant déjà partie de l'échantillon. Les disparitions n'étaient retirées que si elles étaient désignées par une source indépendante ou après une période prédéfinie.

Les mises à jour de l'échantillon, selon les méthodes de Kish et Scott (1971), visaient à maximiser le chevauchement entre les échantillons actuels et les nouveaux échantillons, afin de minimiser les perturbations, les coûts d'exploitation et les discontinuités dans les estimations. Des détails sur la procédure par panel sont présentés dans Hidiroglou, Choudhry et Lavallée (1991).

4.3 Scénarios de renouvellement indépendant du plan d'enquête

Un élément clé de la coordination des échantillons est l'utilisation d'un nombre aléatoire permanent (NAP). Le NAP est un nombre aléatoire uniformément distribué dans l'intervalle [0, 1], attribué de façon permanente à chaque unité de la population. Ce nombre reste attaché à l'unité tant qu'elle existe et constitue une référence continue et identifiable dans la base de population.

L'échantillonnage fondé sur les NAP a été élaboré indépendamment au début des années 1970 par le Australian Bureau of Statistics et Statistics Sweden. Il permet de concevoir des plans de sondage flexibles et coordonnés dans le temps et entre enquêtes, grâce à une méthode cohérente et reproductible de sélection des unités.

Trois procédures d'échantillonnage couramment utilisées et fondées sur les NAP sont :

- *L'échantillonnage de Poisson* : les unités sont sélectionnées indépendamment, avec des probabilités d'inclusion définies en fonction du NAP.
- *L'échantillonnage synchronisé* : il garantit l'inclusion ou l'exclusion cohérente des unités dans l'ensemble des enquêtes coordonnées en alignant les intervalles de sélection sur le NAP.
- *Échantillonnage par microstrates* : il divise l'intervalle du NAP en petites strates (microstrates) pour obtenir une sélection et un renouvellement contrôlés à l'intérieur des strates.

4.3.1 Échantillonnage de Poisson

Une première application du NAP a été l'échantillonnage de Poisson. Brewer, Early et Joyce (1972) ont proposé la méthode suivante pour sélectionner un échantillon de Poisson, comme moyen de gérer à la fois le renouvellement et le chevauchement.

Pour une population U de taille N, on attribue à la k^{e} unité un nombre aléatoire $r_k, k = 1,\ldots,N$ distribué uniformément dans l'intervalle [0,1]. La k^{e} unité est incluse dans l'échantillon initial si : $r_k \in [0, \pi_k]$ où π_k est la probabilité d'inclusion dans l'échantillon, définie par $\pi_k = n x_k \Big/ \sum_{j=1}^{N} x_j$, où x_j est la mesure de taille associée à la j^{e} unité dans la population U. La taille de l'échantillon réalisée n_s est une variable aléatoire avec une moyenne de $E(n_s) = \sum_{k=1}^{N} \pi_k$ et une variance de $V(n) = \sum_{k=1}^{N} \pi_k (1 - \pi_k)$. Le renouvellement de l'échantillon s'effectue en décalant progressivement la fenêtre d'échantillonnage d'un décalage τ vers la droite. Ainsi, lors de la deuxième sélection, la k^{e} unité est incluse dans l'échantillon si (i.) $r_k \in [\tau, \pi_k + \tau]$ ou (ii.) $r_k \in [\tau, 1] \cup [0, \pi_k + \tau - 1]$, où $\pi_k + \tau > 1$. L'inconvénient principal de cette méthode est que la taille de l'échantillon réalisée est une variable aléatoire pouvant présenter une variance relativement élevée.

4.3.2 Échantillonnage synchronisé

La méthode de sélection a été adaptée de la méthode JALES (Atmer, Thulin et Bäcklund, 1975). JALES correspond aux initiales des auteurs qui l'ont développée : Johan Almer et Lars-Erik Sjoberg, tous deux anciens employés de Statistics Sweden. Une bonne description de cette méthode est donnée dans Brewer, Gross et Lee (2000) et McKenzie et Gross (2000). L'Australian Bureau of Statistics a adopté l'échantillonnage synchronisé au début des années 1980 (voir Hinde et Young, 1984 pour une description détaillée de son application informatique). Cette méthode permet de contrôler le renouvellement au niveau de l'enquête, d'autoriser différentes stratifications dans les enquêtes, et de contrôler le chevauchement (coordination) des unités échantillonnées entre les enquêtes. La coordination peut être positive ou négative. Coordination négative : le chevauchement attendu est minimisé. Coordination positive : le chevauchement de l'échantillon attendu est maximisé.

Le renouvellement s'effectue de la manière suivante. Pour la première sélection, un point de départ provisoire s_0, est établi, soit de manière délibérée, soit de façon aléatoire, dans l'intervalle [0, 1]. Étant donné la taille d'échantillon requise n, les premières unités n situées à la position de s_0 ou à sa droite sont sélectionnées. Ces unités sont définies par un intervalle de sélection $[s_1, e_1)$ fermé à gauche et ouvert à droite, où le point de départ s_1 est défini comme la première unité située à la position de s_0 ou à sa droite, et l'extrémité, e_1, est défini comme la $(n+1)^e$ unité à la position de s_0 ou à sa droite. Pour la sélection suivante, il suffit de se rappeler du point de départ et des extrémités s_1 et e_1, et le nombre d'unités de cette nouvelle population dans l'intervalle $[s_1, e_1)$ est comparée avec la taille d'échantillon désirée n. Si le bon nombre d'unités se trouve dans l'intervalle (c'est-à-dire il n'y a eu ni aucune création ni disparition nettes dans l'intervalle de sélection), alors les nouveaux points de départ et extrémité s_2 et e_2, restent généralement s_1 et e_1. Une exception survient si l'unité associée à s_1 ou e_1 disparaît, auquel cas, la prochaine unité à droite devient le point limite. Le chevauchement entre les enquêtes est minimisé de la façon suivante : lorsque le renouvellement a lieu, la position de l'intervalle de sélection change d'un cycle de l'enquête à un autre. Pour contrôler ce chevauchement entre les enquêtes, chaque strate de chaque enquête reçoit un intervalle fixe, appelé étendue de l'enquête, à l'intérieur de laquelle l'intervalle de sélection est contraint de rester.

L'étendue de l'enquête doit être suffisamment large pour contenir l'intervalle de sélection, tout en incluant un nombre suffisant d'unités non sélectionnées afin de permettre un renouvellement efficace. Le chevauchement entre les enquêtes est maximisé de la façon suivante : si deux enquêtes utilisent la même stratification, le chevauchement entre elles est maximisé en attribuant un intervalle de sélection à l'enquête la plus petite et un intervalle disjoint à l'échantillon supplémentaire de la plus grande enquête. Lorsque les enquêtes ont des stratifications différentes, une coordination positive n'est possible que par un choix manuel de l'intervalle de sélection.

4.3.3 Microstrates

La méthode des microstrates, décrite en détail par Rivière (2001), constitue une approche robuste de la coordination des échantillons. Elle est conçue pour gérer le fardeau de réponse et optimiser le contrôle du

chevauchement dans les enquêtes-entreprises. Cette méthode repose sur les permutations systématiques des nombres aléatoires initialement attribués au sein de sous-groupes soigneusement construits de la population, appelés *microstrates*. Ces permutations sont utilisées de manière stratégique pour répondre à plusieurs objectifs de coordination, notamment l'étalement du fardeau de réponse au fil du temps, la réduction du chevauchement des échantillons entre les enquêtes et le maintien de panels à jour.

Le concept de microstratification a été élaboré en 1998 sous l'égide du projet SUPCOM d'Eurostat (Rivière, 1998) sur la coordination des échantillons, avec une première mise en œuvre du logiciel réalisée dans le système SALOMON en 1999. Des améliorations méthodologiques supplémentaires ont abouti à la diffusion de MICROSTRAT en 2001, qui offrait des fonctionnalités élargies et une plus grande flexibilité pour la mise en application pratique. Une bonne discussion de cette procédure se trouve dans Nedyalkova, Pea et Tillé (2008).

Selon cette méthode, chaque unité de la base de sondage (par exemple le RE) comprise dans la portée se voit attribuer un nombre aléatoire fixe tiré d'une répartition uniforme. Ce nombre sert de base à la sélection de l'échantillon. De plus, chaque unité reçoit un coefficient de fardeau de réponse, qui s'accumule au fil du temps en fonction de son inclusion dans une ou plusieurs enquêtes. Bien que les numéros aléatoires eux-mêmes demeurent statiques, des permutations sont autorisées entre les unités, à condition qu'elles soient liées à leurs coefficients respectifs de fardeau de réponse. Ces permutations s'effectuent au sein des *microstrates*, lesquelles sont définies comme le classement recoupé le plus fin possible des unités, permettant un tri selon un fardeau de réponse croissant sans introduire de biais de sélection.

Les principales propriétés statistiques de la méthode des microstrates comprennent :

- *Préservation du caractère aléatoire* : les nombres aléatoires fixes conservent leur répartition uniforme après permutation, ce qui garantit la validité des probabilités de sélection.
- *Prévention des biais* : les permutations à l'intérieur des microstrates sont établies de façon à maintenir l'absence de biais du plan de sondage dans les estimations d'enquête.
- *Traitement automatique de la dynamique* : la méthode intègre aisément les créations d'unités, les disparitions d'unités, et les changements de strates, offrant un haut degré d'adaptabilité aux mises à jour de la base de sondage et aux évolutions de la population.
- *Souplesse de la stratification* : il n'existe pas de contraintes strictes quant au nombre de strates ou aux taux de renouvellement, ce qui accroît son applicabilité générale à diverses enquêtes-entreprises.

Malgré ses avantages, une limite notable réside dans la possibilité que se forment des microstrates trop petites, ce qui peut réduire l'efficacité de la coordination. Toutefois, ce problème peut être atténué par des méthodes alternatives de tri ou des règles d'agrégation qui préservent les objectifs de coordination sans compromettre l'intégrité statistique des strates.

Au Canada, l'EERH a adopté la méthode des microstrates pour ses procédures de renouvellement d'échantillon à partir de 2010, comme le documente Landry (2011).

4.3.4 Comparaison entre le scénario de renouvellement indépendant du plan d'enquête et le scénario de renouvellement dépendant du plan d'enquête

Srinath et Carpenter (1995) ont mené une étude comparative de plusieurs scénarios de renouvellement sur cinquante périodes consécutives, en évaluant à la fois l'échantillonnage par panel (un scénario de renouvellement dépendant du plan d'enquête) et l'échantillonnage de Bernoulli (un scénario de renouvellement indépendant du plan d'enquête). Leur modèle de simulation tenait compte des changements dynamiques de la population en intégrant *les créations* (nouvelles unités entrant dans la population) et les disparitions (unités sortant de la population) pour chaque période, ce qui offrait une évaluation réaliste de la performance des méthodes de renouvellement au fil du temps.

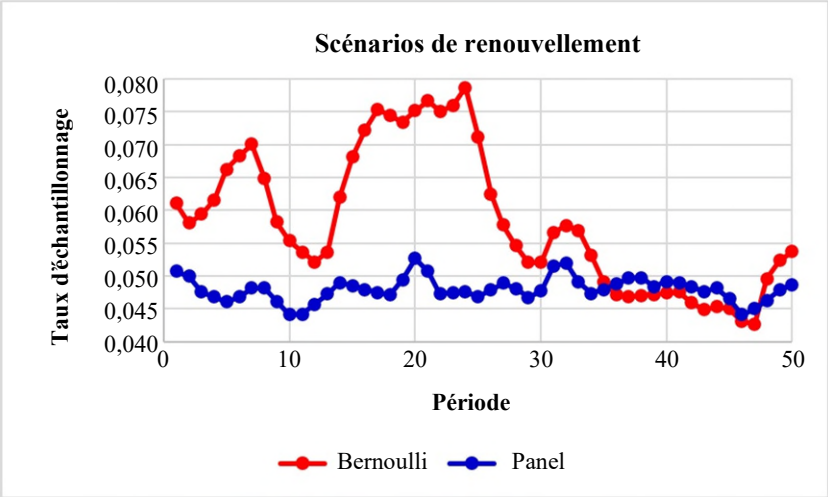
L'échantillon initial a été tiré d'une population de taille 500. À chaque cycle subséquent, le nombre de créations et de disparitions était censé suivre les lois de Poisson selon des moyennes respectives de 20 et de 15. De plus, on supposait une probabilité de 70 % de déterminer une disparition par une source externe.

Les résultats sont illustrés à la figure 4.2, qui correspond à la figure 10.3 (A) de Srinath et Carpenter (1995). Cette figure montre que l'échantillonnage de Bernoulli présente la plus grande variation des fractions d'échantillonnage au fil du temps, ce qui met en évidence l'une de ses principales limites pour le maintien de la stabilité de l'échantillon.

Les différences importantes dans les taux d'échantillonnage attendus entre l'échantillonnage par panel et l'échantillonnage de Bernoulli ont des répercussions importantes pour l'estimation des totaux. L'échantillonnage par panel tend à produire une série chronologique plus régulière que l'échantillonnage de Bernoulli, à condition que les créations et les disparitions ne perturbent pas fortement l'équilibre du nombre d'unités entre les panels. Toutefois, si un déséquilibre survient dans le nombre d'unités entre les panels, même l'estimation composite peut ne pas parvenir à fournir des estimations stables d'une édition de l'enquête à une autre.

L'Enquête mensuelle sur le commerce de gros et de détail, remaniée au milieu des années 1980, comprenait à l'origine un plan d'échantillonnage par panel. Toutefois, le renouvellement des unités a été interrompu tôt dans sa mise en œuvre, car les estimations d'un mois à l'autre présentaient une variabilité excessive. Cette instabilité était principalement attribuable à une stratification inadéquate des petites unités dans la base de sondage.

Figure 4.2 Comparaison des taux d'échantillonnage de deux scénarios de renouvellement



En revanche, les enquêtes mensuelles sur le commerce de détail et de gros du U.S. Census Bureau utilisaient un plan de sondage dans lequel les entreprises échantillonnées étaient réparties en trois panels, chacune répondant tous les trois mois. Cette approche permettait de disposer de données qui se chevauchaient pour le mois courant et le mois précédent, et l'estimation composite était utilisée pour réduire la variance tant des niveaux que des changements. Malgré ces efforts, le renouvellement des échantillons pour ces enquêtes a été abandonné en 1997 au profit d'un plan sans panel (voir Cantwell et Caldwell, 1998 pour plus de détails).

Ces enquêtes publient des estimations de totaux. L'EERH qui repose sur le renouvellement d'échantillon depuis le début des années 1980 et continue de le faire selon la méthode MICROSTRAT, présente des changements d'estimations entre périodes successives, telles que la rémunération hebdomadaire moyenne et l'emploi salarié, qui demeurent dans des limites acceptables. L'une des raisons est que ces variables présentent beaucoup moins d'asymétrie que celles du commerce de détail et de gros, ce qui conduit à des estimations plus stables et fiables.

5. Élimination des unités disparues

Les disparitions s'accumulent au fil du temps dans une enquête périodique, en particulier lorsqu'elle est réalisée fréquemment (par exemple chaque mois ou chaque trimestre) et que l'échantillon n'est pas sélectionné de manière indépendante à chaque cycle. Bien que le maintien des unités disparues dans l'échantillon n'entraîne pas de biais dans les estimations, il provoque une hausse de la variance au fil du temps, puisque ces unités sont codées à zéro. Pour atténuer cet effet, il est important de retirer les unités disparues à la fois de l'échantillon et de la base de population. Toutefois, ce retrait doit être réalisé avec précaution pour éviter d'introduire un biais dans les estimations au fil du temps. Les poids d'échantillonnage initiaux, fondés sur

les fractions d'échantillonnage d'origine, doivent être ajustés afin de tenir compte des répercussions du retrait des unités disparues.

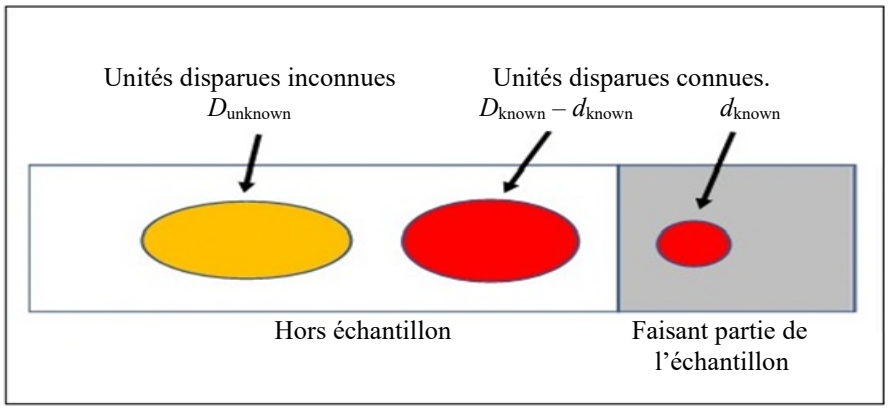
Ce qui suit décrit la procédure au niveau de la strate secondaire U, définie selon la taille des unités. Supposons qu'à un moment donné de l'enquête, la strate U compte N unités de population, dont n unités sont sélectionnées par échantillonnage aléatoire simple sans remplacement. Le taux d'échantillonnage obtenu est $f = n/N$, et le poids d'échantillonnage initial associé à chaque unité échantillonnée est $w = 1/f$. Ce poids demeure fixe pendant toute la durée de vie de l'enquête.

Supposons que D_{known} et d_{known} représentent respectivement le nombre *connu* d'unités disparues dans la strate U et le nombre *connu* d'unités disparues dans l'échantillon s. Alors la différence, $D_{\text{known}} - d_{\text{known}}$, correspond aux unités disparues *connues* qui se trouvent hors de l'échantillon $U \setminus s$. Ces unités disparues *connues* sont déterminées soit par l'échantillon de l'enquête, soit par une source administrative qui met à jour le registre de la population. En revanche, les unités disparues *inconnues* désignent les unités hors échantillon qui sont disparues, mais qui n'ont pas encore été déterminées par la source administrative. Supposons que D_{unknown} désigne le nombre d'unités disparues *inconnues* hors échantillon.

La figure 5.1 illustre le problème : comment déterminer le nombre d'unités disparues à conserver dans l'échantillon et hors échantillon afin qu'elles représentent fidèlement le nombre inconnu d'unités disparues hors échantillon?

L'estimation du nombre total d'unités disparues connues et inconnues dans la population est wd_{known}. Cela suppose que l'estimation du nombre d'unités connues et inconnues hors échantillon est $(w-1)\,d_{\text{known}}$. Il s'ensuit que l'estimation du nombre d'unités inconnues hors échantillon est $\hat{D}_{\text{unknown}} = (w-1)\,d_{\text{known}} - (D_{\text{known}} - d_{\text{known}})$. Cette différence peut être négative, nulle ou positive. Si $\hat{D}_{\text{unknown}} \leq 0$, aucune unité disparue (ni faisant partie de l'échantillon l'échantillon ni hors échantillon) n'est retirée. Si $\hat{D}_{\text{unknown}} > 0$, on peut retirer certaines unités disparues *connues* faisant partie de l'échantillon, ainsi que certaines unités disparues *connues* hors échantillon, de sorte que le nombre d'unités disparues restantes faisant partie de l'échantillon représente les unités hors échantillon disparues *inconnues*.

Figure 5.1 Disparitions connues ou inconnues



Supposons que x est le nombre d'unités disparues à conserver dans l'échantillon. Ces unités représentent $(w-1)\,x = \hat{D}_{\text{unknown}}$ *disparitions inconnues* dans la population.

Sachant que $\hat{D}_{\text{unknown}} = (w-1)\,d_{\text{known}} - (D_{\text{known}} - d_{\text{known}}),$ nous avons

$$\left(w-1\right)\,x\; = (w-1)\,d_{\text{known}} - (D_{\text{known}} - d_{\text{known}}). \tag{5.1}$$

En résolvant l'équation (5.1) pour $x,$ le nombre d'unités disparues à conserver dans l'échantillon s pour représenter les disparitions hors échantillon inconnues est :

$$x = d_{\text{known}} - d^{*}_{\text{known}}$$

où $d^{*}_{\text{known}} = (D_{\text{known}} - d_{\text{known}}) \,/\, (w-1).$ Dans le cas présent, d^{*}_{known} est le nombre d'unités disparues à retirer de l'échantillon.

De façon correspondante, $w\,d^{*}_{\text{known}}$ estime le nombre total de disparitions retirées des d_{known} unités disparues dans l'échantillon. Ces disparitions ne peuvent pas être retirées des unités disparues inconnues $D_{\text{unknown}},$ elles doivent donc être retirées des unités disparues connues. L'estimation du nombre d'unités disparues connues est $\hat{D}^{*}_{\text{known}} = w\,d^{*}_{\text{known}}.$ Il s'ensuit que le nombre d'unités disparues qui devraient être retirées des unités hors échantillon $(D_{\text{known}} - d_{\text{known}})$ devrait être $\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}}.$ Il convient de mentionner que le nombre d'unités à retirer de l'échantillon (d^{*}_{known}) et hors échantillon $(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})$ sera en général un nombre réel; par conséquent, il doit être converti en nombre entier. À cette fin, on note la partie entière de d^{*}_{known} par $d^{*}_{\text{known, int}},$ où $d^{*}_{\text{known, int}} \le d^{*}_{\text{known}}.$ Ainsi, le nombre d'unités à retirer des unités disparues faisant partie de l'échantillon sera :

$$d^{*}_{\text{known, int}} = \text{int}\left\{\min\left(d^{*}_{\text{known}}, d_{\text{known}}\right)\right\} \tag{5.2}$$

où int (a) désigne la partie entière du nombre réel $a.$

Le nombre d'unités disparues à conserver dans l'échantillon est $d_{\text{known}} - d^{*}_{\text{known, int}}.$ De façon correspondante, le nombre d'unités disparues connues hors échantillon qui doivent être retirées des unités disparues $D_{\text{known}} - d_{\text{known}}$ hors échantillon est :

$$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}} = \text{int}\left\{(w-1)\,d^{*}_{\text{known}} + 0{,}5\right\}. \tag{5.3}$$

Cela conduit à trois scénarios pour le retrait combiné des unités disparues faisant partie de l'échantillon et hors échantillon, résumés au tableau 5.1

Tableau 5.1
Résultats du retrait des unités disparues

Cas	Condition	Unités disparues retirées faisant partie de l'échantillon	Unités disparues retirées hors échantillon
1.	$d^{*}_{\text{known, int}} = 0$	Aucune unité disparue	Aucune unité disparue retirée
2.	$d^{*}_{\text{known, int}} = d_{\text{known}}$	Toutes les unités disparues	$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}}$ unités disparues retirées
3.	$0 < d^{*}_{\text{known, int}} < d_{\text{known}}$	$d^{*}_{\text{known, int}}$ unités disparues	$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}}$ unités disparues retirées

Quand les unités disparues sont retirées de l'échantillon et hors échantillon, il reste $(N - d^*_{\text{known,int}}) - (\hat{D}^*_{\text{known}} - d^*_{\text{known}})_{\text{int}}$ unités dans la population de la strate et $n - d^*_{\text{known,int}}$ unités dans l'échantillon. Le retrait des unités disparues devrait entraîner seulement de légères différences entre le poids d'échantillonnage précédent w et le nouveau poids.

$$w^* = \left(N - d^*_{\text{known,int}} - (\hat{D}^*_{\text{known}} - d^*_{\text{known}})_{\text{int}}\right) \Big/ (n - d^*_{\text{known,int}}).$$

Désignons respectivement U_{c} l'univers et s_c l'échantillon après le retrait des unités disparues. Si l'estimateur est un total fondé sur une variable d'intérêt y, on pourrait estimer le total de la population par $\hat{Y}^*_{\text{HT}} = \sum_{i \in s_{\text{c}}} w y_i$.

Cependant, un estimateur amélioré, qui reflète le fait que le retrait des unités disparues a eu des répercussions à la fois sur le compte de la population et sur le compte de l'échantillon, est l'estimateur de Hájek, obtenu par :

$$\hat{Y}^*_{\text{HAJ}} = \frac{\left(N - d^*_{\text{known,int}} - (\hat{D}^*_{\text{known}} - d^*_{\text{known}})_{\text{int}}\right)}{\sum_{i \in s_c} w^*} \sum_{i \in s_c} w^* y_i.$$

6. Résumé

Les enquêtes-entreprises jouent un rôle crucial dans la collecte de données qui appuient la prise de décision éclairée, tant dans le secteur public que privé. La qualité de ces enquêtes repose sur plusieurs facteurs clés, notamment : i) un RE complet, non dupliqué et à jour; ii) un plan de sondage qui reflète la structure réelle des entreprises et permet une sélection optimale de l'échantillon; iii) l'intégration à des sources de données administratives pour accroître l'efficacité et rehausser la qualité des données; iv) un système de traitement générique qui soutient toutes les étapes des opérations d'enquête-entreprises, y compris la collecte des données, la vérification, l'estimation, l'échantillonnage et l'imputation; v) des mécanismes de rétroaction efficace entre le RE et les enquêtes qui en sont tirées; vi) des stratégies visant à réduire le fardeau de réponse, telle que des questionnaires bien conçus, le renouvellement des échantillons dans une enquête donnée et la coordination entre les enquêtes pour limiter la sélection répétée des mêmes entreprises; vii) le retrait des unités disparues qui assure une estimation correcte de la variance.

La présente étude a permis de se concentrer principalement sur : i) l'établissement, la structure et la mise à jour continue du registre des entreprises; ii) les méthodes de détermination et de répartition de la taille d'échantillon; iii) les techniques de renouvellement et de coordination des échantillons; iv) une procédure visant l'élimination des unités disparues.

Ces améliorations méthodologiques ont évolué et ont permis à Statistique Canada de mettre en place un programme de la statistique des entreprises robuste et fiable.

Remerciements

L'auteur remercie l'éditeur et les examinateurs pour leurs excellents commentaires, qui ont permis d'améliorer cet article.

Annexe A

Extension multivariée de Bankier (1988)

Définissons un vecteur multivarié $\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kJ})$ pour $k = 1, \ldots, N,$ de dimension J. Le problème consiste à répartir une taille d'échantillon n sélectionnée dans une population U de taille N entre H strates $(U_h, h = 1, \ldots, H)$. En prolongeant l'approche de Bankier (1988), on minimise la fonction F_2

$$F_2 = \sum_{h=1}^{H}\sum_{j=1}^{J}\left(X_h^q r_j \mathrm{CV}(\hat{Y}_{hj})\right)^2 = \sum_{h=1}^{H}\sum_{j=1}^{J}\left(X_h^{2q} r_j^2 \frac{V(\hat{Y}_{hj})}{Y_{hj}^2}\right),$$

sous la contrainte $\sum_{h=1}^{H} n_h \kappa_h = \mathrm{K}, h = 1, \ldots, H,$ où

- κ_h est le coût d'énumération par unité dans la strate h;
- n_h est la taille de l'échantillon alloué à la strate h;
- K est le coût total alloué;
- X_h^q est l'importance de la strate $h,$ et q est une constante dans l'intervalle $0 \le q \le 1$;
- r_j est l'importance relative de la variable y_j indépendamment de la strate;
- $Y_{hj} = \sum_{k \in U_h} y_{kj}$ est le total de la variable y_j dans la strate h;
- $\hat{Y}_{hj} = N_h \sum_{k \in s_h} y_{kj} / n_h$ fait l'estimation du total Y_{hj}.

La variance $V(\hat{Y}_{hj})$ est

$$V(\hat{Y}_{hj}) = N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{hj}^2;\ h = 1, 2, \ldots, H, j = 1, 2, \ldots, J,$$

et $S_{hj}^2 = \sum_{k \in U_h} (y_{kj} - \bar{Y}_{hj})^2 \big/ (N_h - 1).$

La fonction F_2 est minimisée à l'aide de la fonction de Lagrange $F_3,$ où

$$F_3 = \sum_{h=1}^{H}\sum_{j=1}^{J}\left(X_h^{2q} r_j^2 A_{hj}^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\right) + \lambda\left(\sum_{h=1}^{H} n_h \kappa_h - K\right),$$

avec $A_{hj}^2 = N_h^2 S_{hj}^2 / Y_{hj}^2$ et $\lambda,$ qui est le multiplicateur de Lagrange. F_3 est différenciée par rapport à n_h $(h = 1, \ldots, H)$ et λ. En établissant les équations obtenues à zéro, on obtient

$$\frac{\partial F_3}{\partial n_h} = -\sum_{j=1}^{J}\left(X_h^{2q} r_j^2 \frac{A_{hj}^2}{n_h^2}\right) + \lambda \kappa_h = 0, \tag{A.1}$$

et

$$\frac{\partial F_3}{\partial \lambda} = \sum_{h=1}^{H} n_h \kappa_h - K = 0. \tag{A.2}$$

En résolvant n_h à partir de l'équation (A.1), on obtient :

$$n_h = \frac{\sqrt{\sum_{j=1}^{J} X_h^{2q} r_j^2 A_{hj}^2 / \kappa_h}}{\sqrt{\lambda}}. \tag{A.3}$$

En substituant n_h de (A.3) dans (A.2), on obtient :

$$\sum_h n_h \kappa_h - K = \sum_{h=1}^{H} \frac{\sqrt{\kappa_h \sum_{j=1}^{J} X_h^{2q} r_j^2 A_{hj}^2}}{\sqrt{\lambda}} - K = 0. \tag{A.4}$$

D'après l'équation (A.4), nous avons :

$$\frac{1}{\sqrt{\lambda}} = \frac{K}{\sum_{h=1}^{H} \sqrt{\kappa_h \sum_{j=1}^{J} X_h^{2q} r_j^2 A_{hj}^2}}. \tag{A.5}$$

En substituant $1/\sqrt{\lambda}$ de (A.5) dans (A.3), on obtient :

$$n_h = K \frac{\sqrt{\sum_{j=1}^{J} (X_h^{2q} r_j^2 A_{hj}^2)/\kappa_h}}{\sum_{h=1}^{H} \sqrt{\sum_{j=1}^{J} (X_h^{2q} r_j^2 A_{hj}^2)/\kappa_h}}.$$

Si $J = 1$, nous obtenons la version unidimensionnelle comme suit :

$$n_h = K \frac{\sqrt{(X_h^{2q} A_h^2)/\kappa_h}}{\sum_{h=1}^{H} \sqrt{(X_h^{2q} A_h^2)/\kappa_h}}. \tag{A.6}$$

Si $\kappa_h = \kappa$ pour $h = 1, \ldots, H$, $K = n$, alors n_h est

$$n_h = n \frac{S_h X_h^q / \bar{Y}_h}{\sum_{h=1}^{H} S_h X_h^q / \bar{Y}_h}.$$

Cela correspond à l'équation (3.2) : les termes gi dans cette équation sont remplacés par h.

Annexe B

Équivalence entre les scénarios c et n

Supposons que la version continue de la répartition discrète de la variable x est $g(x)$. La version continue de l'équation (3.8), utilisée pour minimiser la taille d'échantillon global n pour un coefficient de variation déterminé c est donnée par :

$$n = N \omega_H + \frac{N\ AB}{F} \tag{B.1}$$

où

$$A = \sum_{h=1}^{H-1} \omega_h^{2q_1} \mu_h^{2q_2} \sigma_h^{2q_3},\ B = \sum_{h=1}^{H-1} (\omega_h \sigma_h)^2\ \omega_h^{-2q_1} \mu_h^{-2q_2} \sigma_h^{-2q_3},\ D = \sum_{h=1}^{H-1} \omega_h \sigma_h^2$$

et $F = N(c\mu)^2 + D$.

Les termes ω_h μ_h, et μ dans A, B, D, et F, sont respectivement les analogues continus de W_h $\bar{X}_h$ σ_h^2, et de $\bar{X}$. En faisant la différentiation de (B.1) par rapport à $b_h, h = 1, 2, \ldots, H-1$, et en faisant correspondre les équations qui en découlent égales à 0, on obtient :

$$F(A_h^{(1)} B + A B_h^{(1)}) - A B D_h^{(1)} = 0 \text{ pour } h = 1, \ldots, H-2 \tag{B.2}$$

et

$$F\,(A_h^{(1)}\,B + A\,B_h^{(1)}) - A\,B\,D_h^{(1)} + F^2\,\omega_L^{(1)} = 0 \text{ pour } h = H-1 \tag{B.3}$$

où $A_h^{(1)}$, $B_h^{(1)}$, et $D_h^{(1)}$ sont les dérivés de $A, B,$ et D par rapport à b_h; $h = H-1,$ et $\omega_H^{(1)}$ est la dérivée de ω_L par rapport à b_h pour $h = H-1.$

La version continue de l'équation (3.7), qui permet de minimiser $V(\hat{\bar{X}})$ pour une taille d'échantillon déterminée n est donnée par :

$$V(\hat{\bar{X}}) = \frac{AB}{n - N\omega_H} - \frac{D}{N} \tag{B.4}$$

En faisant la différentiation de (B.4) par rapport à b_h, $h = 1,2,\ldots,H-1,$ et en faisant correspondre l'équation obtenue égale à zéro, on obtient :

$$\frac{\left(A_h^{(1)}B + AB_h^{(1)}\right)}{n - N\,\omega_L} - \frac{D_h^{(1)}}{N} = 0 \text{ pour } h = 1,...,H-2 \tag{B.5}$$

et

$$\frac{\left(A_h^{(1)}B + AB_h^{(1)}\right)}{n - N\,\omega_L} - \frac{D_h^{(1)}}{N} + \frac{AB}{\left(n - N\,\omega_L\right)^2} N\,\omega_L^{(1)} = 0 \text{ pour } h = H-1. \tag{B.6}$$

Nous démontrons maintenant que les équations (B.2) et (B.5) sont équivalentes pour $h = 1,\ldots,H-2.$ D'après l'équation (B.1), nous avons :

$$n - N\,\omega_L = \frac{N\,AB}{F}. \tag{B.7}$$

En substituant (A.7) dans (A.5), on obtient :

$$\begin{aligned}\frac{(A_h^{(1)}B + AB_h^{(1)})}{n - N\,\omega_L} - \frac{D_h^{(1)}}{N} &= \frac{1}{NAB}\left\{F(A_h^{(1)}B + AB_h^{(1)}) - ABD_h^{(1)}\right\} \\ &= 0.\end{aligned}$$

Par conséquent, le résultat s'ensuit. Par un raisonnement analogue, on peut démontrer que les équations (A.3) et (A.6) sont équivalentes pour $h = H-1.$

Bibliographie

Atmer, J.G., Thulin, G. et Bäcklund, S. (1975). Coordination of samples with the JALES technique. *Statistisk Tidskrift*, 13, 443-450. https://www.yumpu.com/sv/document/view/20050785/1975-statistical-review-third-series-vol-13-1975-statistiska-#.

Baillargeon, S. et Rivest, L.-P. (2009). A general algorithm for univariate stratification. *Revue Internationale de Statistique*, 77, 331-344.

Baillargeon, S., Rivest, L.-P. et Ferland, M. (2007). Stratification en enquêtes entreprises : Une revue et quelques avancées. *Recueil de la Section des méthodes d'enquête,* Société Statistique du Canada. https://ssc.ca/sites/default/files/survey/documents/SSC2007_S_Baillargeon.pdf.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42(3), 174-177. https://doi.org/10.1080/00031305.1988.10475556.

Bérard, H. (2001). The redesign of the monthly wholesale and retail trade survey of Statistics Canada. *Recueil de la Section des méthodes d'enquête*, Société Statistique du Canada, 81-86. https://ssc.ca/sites/default/files/survey/documents/SSC2001_H_Berard.pdf.

Brewer, K.R.W., Early, L.J. et Joyce, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

Brewer, K.R.W., Gross, W.F. et Lee, G.F. (2000). PRN Sampling: The Australian Experience. *Recueil : communications invitées, sujets IASS*, Helsinki du 10 au 18 août 1999, 155-163. https://ww2.amstat.org/meetings/ices/2000/proceedings/S31.pdf.

Brodeur, M., Koumanakos, P., Leduc, J., Rancourt, É. et Wilson, K. (2005). L'approche intégrée des enquêtes économiques au Canada. Statistique Canada, nº 68-514-XWE au catalogue. https://www150.statcan.gc.ca/n1/pub/68-514-x/68-514-x2006001-fra.htm.

Cantwell, P. et Caldwell, C. (1998). Examining the revisions in monthly retail and wholesale trade surveys under a rotating panel design. *Journal of Official Statistics*, 14(1), 47-59.

College, M.J. (1995). Frames and business registers: An overview. Dans *Business Survey Methods*, (Éds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College et P.S. Kott), 1995 Wiley-Interscience Publication Toronto, 21-47.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 1950(3-4), 203-213.

Daoust, P., Mireuta, M. et Kleim, G. (2023). The integrated business statistics program is ten years old! A methodological perspective. *Proceedings of Session CPS 11 - Finance and Business Statistics VII. 64th ISI World Statistics Congress*, juillet 2023. https://www.isi-next.org/media/abstracts/ottawa-2023_8dcef16b7fa9e697566aa54408a57532.pdf.

Deming, W.E. et Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427-444.

EERH (2025). Enquête sur l'emploi, la rémunération et les heures de travail, Information détaillée pour février 2025. https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&SDDS=2612.

EMCD (2025). Enquête mensuelle sur le commerce de détail, Information détaillée pour février 2025. https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&Id=1570101.

EMCG (2025). Enquête mensuelle sur le commerce de gros, Information détaillée pour février 2025. https://www23.statcan.gc.ca/imdb/p2SV_f.pl?Function=getSurvey&Id=1566980.

Ferland, M. et Batten, D. (2007). Sampling for the MWRTS Redesign. Document de travail no BSMD-2007-002E, Direction de la méthodologie, Ottawa : Statistique Canada.

Hidiroglou, M.A. (1986a). Wholesale/Retail Trade Surveys Redesign. Document de travail no BSMD-87-017E, Direction de la méthodologie, Ottawa : Statistique Canada.

Hidiroglou, M.A. (1986b). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40(1), 27-31.

Hidiroglou, M.A., Choudhry, G.H. et Lavallée, P. (1991). Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises. *Techniques d'enquête*, 17(2), 211-227. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14502-fra.pdf.

Hidiroglou, M.A. et Kozak, M. (2018). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *Revue Internationale de Statistique*, 86(1), 87-105.

Hidiroglou, M.A. et Laniel, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys. *Revue Internationale de Statistique*, 69(3), 487-504.

Hidiroglou, M.A. et Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11(4), 397-405.

Hinde, R. et Young, D. (1984). Synchronized Sampling and Overlap Control Manual. Rapport non publié, Australian Bureau of Statistics.

Kirkendall, N.K., White Jr, G.D., Citro, C.F. et Abraham, K.G. (2018). Reengineering the Census Bureau's Annual Economic Surveys. https://nap.nationalacademies.org/catalog/25098/reengineering-the-census-bureaus-annual-economic-surveys.

Kish, L. et Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66(335), 461-470.

Kozak, M. (2004). Optimal stratification using the random search method in agricultural surveys. *Statistics in Transition*, 6(5), 797-806.

Landry, S. (2011). Managing response burden by controlling sample selection and survey coverage. *Proceedings of Survey Research Methods Section*, American Statistical Association, 2276-2284. http://www.asasrms.org/Proceedings/y2011/Files/301419_66701.pdf.

Lavallée, P. et Hidiroglou, M.A. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14(1), 35-45. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1988001/article/14602-fra.pdf.

Lednicki, B. et Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

McKenzie, R. et Gross, B. (2000). Synchronised sampling. *Proceedings of the Second International Conference on Establishment Surveys Methods for Businesses, Farms, and Institutions*, du 17 au 21 juin 2000, 237-254. https://ww2.amstat.org/meetings/ices/2000/proceedings/S31.pdf.

Nedyalkova, D., Pea, J. et Tillé, Y. (2008). Sampling procedures for coordinating stratified samples: Methods based on microstrata. *Revue Internationale de Statistique*, 76(3), 368-386.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28(2), 207-214. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2002002/article/6432-fra.pdf.

Rivest, L.-P. et Baillargeon, S. (2022). *stratification*: *Univariate Stratification of Survey Populations.* https://cran.r-project.org/web/packages/stratification/stratification.pdf.

Rivière, P. (1998). Description of the chosen method: Deliverable 2 of the SUPCOM 1996 project (part "Co-ordination of samples"), EUROSTAT report, 11-33.

Rivière, P. (2001). Coordination d'échantillons par la méthode des microstrates. Recueil: *Symposium 2001, La qualité des données d'un organisme statistique : une perspective méthodologique*. Statistique Canada. https://www150.statcan.gc.ca/n1/fr/pub/11-522-x/2001001/session8/6252-fra.pdf.

Şchiopu-Kratina, I. et Srinath, K.P. (1991). Renouvellement de l'échantillon et estimation dans l'Enquête sur l'emploi, la rémunération et les heures de travail. *Techniques d'enquête*, 17(1), 89-100. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1991001/article/14520-fra.pdf.

Sethi, V.K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5(1), 20-33.

Srinath, K.P. et Carpenter, R.M. (1995). Sampling methods for repeated business surveys. Dans *Business Survey Methods*, (Éds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College et P.S. Kott), 1995 Wiley -Interscience Publication Toronto, 173-183.

Statistique Canada (1997). Projet d'amélioration des statistiques économiques provinciales. https://publications.gc.ca/collections/collection_2017/statcan/11-204/CS11-204-1997-fra.pdf.

St-Louis, G. (2008). The evolution of administrative data use for the Canadian Business Register (BR). Conférence de Shanghai, juin 2008. https://www.stats.gov.cn/english/SpecialTopics/IAOSConference/200811/W020130912531695511742.doc.

The Business Number and Your Canada Revenue Agency Program Accounts. https://www.cchwebsites.com/content/pdf/tax_forms/ca/en/rc2_en.pdf.

Trépanier, J., Babyak, C., Marchand, I., Bissonnette, J. et St-Pierre, M. (1998). Enhancements to the Canadian Monthly Wholesale and Retail Trade Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 487-492. http://www.asasrms.org/Proceedings/papers/1998_082.pdf.

Trépanier, J. (2004). The redesigned Canadian Monthly Wholesale and Retail Trade Survey: A post-mortem of the implementation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4515-4522. http://www.asasrms.org/Proceedings/y2004/files/Jsm2004-000775.pdf.