Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Sampling for business surveys at Statistics Canada

by M.A. Hidiroglou

Release date: December 23, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Survey Methodology, December 2025
Vol. 51, No. 2, pp. 341-379
Statistics Canada, Catalogue No. 12-001-X

Sampling for business surveys at Statistics Canada

M.A. Hidiroglou[1]

Abstract

This article examines the methodological complexities associated with the design of business surveys, with particular emphasis on sampling strategies implemented by National Statistical Offices (NSOs). It addresses the inherent challenges posed by the dynamic nature of the business population, which necessitates continual updates to the sampling frame to ensure representativeness and relevance. Critical design considerations include the determination of optimal sample sizes, stratification across key dimensions such as industry, geographic region, and enterprise size, as well as the treatment of business births and the exclusion of inactive (or "dead") units. The article applies Bankier's (1988) power allocation method to a two-way stratification scheme defined by industry and geography, evaluating its performance by comparing the resulting coefficients of variation with those obtained via a raking algorithm applied to the marginal coefficients. Furthermore, the approach is extended to a multivariate context to accommodate multiple estimation domains. The discussion also encompasses practical issues related to sample rotation and coordination, which are critical for maintaining data quality and minimizing respondent burden over time.

Key Words: Bankier method; Business register; Dead unit removal; Microstrata; Neyman allocation; Permanent random numbers; Power allocation; Rotation; Sample coordination.

1. Introduction

Business surveys are typically conducted on a monthly, quarterly, or annual basis to produce estimates of totals, averages, ratios, and temporal changes across a range of key economic and structural indicators. The methodological challenges associated with the design of business surveys are broadly shared among National Statistical Offices (NSOs) and are typically more complex than those encountered in household or social surveys, largely due to the highly dynamic and heterogeneous nature of the business population.

These surveys are essential tools for collecting data from enterprises for diverse statistical and policy purposes. NSOs rely on them to compile critical macroeconomic indicators, including gross domestic product (GDP), investment, government expenditure, and foreign trade statistics, across various reference periods. The scope of data collected through business surveys is extensive and may encompass domains such as production (e.g., outputs, inputs, transportation, environmental emissions), sales (wholesale and retail), commodity flows (e.g., inventories, shipments, new orders), financial performance (e.g., revenues, expenses, assets, liabilities), labor (e.g., employment levels, payroll data, employee demographics), and prices (e.g., producer and consumer price indices).

Given the considerable heterogeneity of businesses in terms of size and industrial classification, careful sample designs are imperative. Business surveys typically draw samples from business registers, which are derived from administrative sources and include essential contact information such as enterprise names, addresses, and business-specific communication details. The purpose of this article is to describe and provide solutions to the sampling of businesses. The sampling method selected will affect the survey's

1. M.A. Hidiroglou, Statistics Canada Alumnus, Ottawa, Ontario. E-mail: hidirog@yahoo.ca.

efficiency and precision. Sample size plays a crucial role in determining the reliability of survey results. A sample that is too small may lead to unreliable estimates, while an excessively large sample can incur unnecessary costs and operational and respondent burden. Survey designers must carefully calculate the appropriate sample size based on the desired level of precision and the variability in the target population, balancing statistical rigor with practical constraints. The target population (also sometimes referred to as the *scope* of the survey) may be all units within the country, within a particular industrial sector, or in a particular region.

The article is organized as follows. Section 2 provides background on the evolution of Statistics Canada's Business Register (BR) since its inception in the 1970s, with a focus on its current architecture as it pertains to sampling applications. Section 3 outlines the methodological framework for stratifying the statistical units identified on the BR across three key dimensions: industry, geography, and enterprise size. Section 4 reviews procedures for initial sample selection, sample rotation, and the inclusion of newly established businesses (births) in ongoing business surveys. These methods are designed to ensure that the sample remains representative of the evolving population over time, thereby supporting the production of accurate and reliable estimates. Section 5 addresses the identification and removal of known inactive units (dead units) both within and outside the sample. A key challenge arises from the disproportionate discovery of known dead units within the sample compared to those outside it. Additionally, a portion of the population outside the sample may consist of businesses that are inactive but have not yet been formally identified as such, referred to as unknown deaths. Known dead units within the sample are typically identified through direct respondent feedback or auxiliary sources, while those outside the sample can only be recognized via independent sources. This asymmetry underscores the need for robust methodologies to manage the removal of inactive units across the entire survey population. A summary is presented in Section 6.

2. Business register

2.1 Some history

In the early 1970s, Statistics Canada established its BR to serve as a foundational frame for a variety of economic surveys. Initially, the BR was constructed primarily using payroll deduction data from Revenue Canada for small businesses and survey feedback for larger enterprises. However, over time, it became apparent that the BR did not adequately meet the evolving needs of economic survey programs, particularly with respect to coverage, timeliness, and analytical tools. One of the primary deficiencies was incomplete coverage; the BR lacked centralized control over necessary data and was fragmented across various divisions and databases. As a result, many major surveys opted not to transition to the BR and instead maintained independent frames, citing the BR's limitations for their specific requirements.

In response to these issues, Statistics Canada launched the Business Survey Redesign Project in 1984. This was a comprehensive initiative aimed at modernizing the agency's infrastructure for economic survey operations. The project's primary objective was to unify and optimize the maintenance of business structures

across surveys. At the time, existing systems were disparate and inefficient, spread across multiple divisions and databases. The redesign sought to consolidate these systems into a cohesive framework that would facilitate improved data consistency, streamline access to business information, and enhance operational efficiency. A central focus was the increased use of administrative data sources, such as income tax records and other government-held business information, to reduce redundancy and improve data quality.

The redesigned survey infrastructure became operational in March 1988, beginning with the Monthly Wholesale and Retail Trade Surveys (MWRTS). These surveys were ideal for early implementation due to their dependence on a current and reliable business frame for collecting wholesale and retail trade data. While both surveys shared a common modular system architecture – with components for sample selection, data collection, editing, imputation, and estimation – they differed in specific methodological details. For example, the Retail Trade Survey targeted locations, whereas the Wholesale Trade Survey focused on establishments. They also varied in the variables collected, their imputation, and estimation procedures.

The modular design was intentionally built for flexibility, allowing it to be adapted and reused across a wide range of business surveys. This marked the beginning of a generalized systems approach to business survey methodology. Detailed documentation of this framework is available in Hidiroglou (1986a). The sample design of the MWRTS underwent several improvements over time, as described in Trépanier, Babyak, Marchand, Bissonnette and St-Pierre (1998), Bérard (2001), Trépanier (2004), and Ferland and Batten (2007).

The successful implementation of the MWRTS marked a turning point in modernizing the business survey infrastructure at Statistics Canada. Building on this success, the associated methodology and processing systems were progressively expanded to encompass a broader range of businesses, including those in manufacturing, services, and other sectors. This phased rollout promoted greater consistency, coherence, and cost-efficiency across survey programs.

A pivotal advancement occurred in 1994 with the introduction of the nine-digit Business Number (BN) by the Canada Revenue Agency (CRA), aimed at streamlining business identification for tax and regulatory purposes (St-Louis, 2008). The BN root consists of the first nine digits of a business's 15-digit Business Number (BN). It is shared by all businesses within the same enterprise and allows the Canada Revenue Agency (CRA) to easily identify related entities. For more information on how the BN is assigned, refer to the CRA document "The Business Number and your Canada Revenue Agency Program Accounts".

The integration of the BN into Statistics Canada's BR in 1997 marked a major improvement in the agency's ability to link administrative datasets. This integration enhanced the accuracy, completeness, and granularity of provincial economic statistics by enabling consistent linkage across diverse data sources, such as tax records and labour surveys. The adoption of a unique business identifier mitigated issues related to inconsistent or incomplete identifiers, thereby improving data quality and reliability.

Concurrently, Statistics Canada initiated a comprehensive redesign of its annual business survey architecture, culminating in the development of the Unified Enterprise Survey (UES) (Brodeur, Koumanakos, Leduc, Rancourt and Wilson, 2005). Conceived as part of the broader Project to Improve Provincial

Economic Statistics (PIPES) (Statistics Canada, 1997), the UES sought to consolidate all annual business surveys into a single, harmonized framework. Its primary objectives were to enhance the consistency, coherence, and comprehensiveness of business statistics across both provincial and industrial dimensions.

In 1998, Statistics Canada undertook a major overhaul of the BR, building on the earlier implementation of the BN and aligning with the objectives of the PIPES initiative. This transformation marked a significant step toward modernizing the agency's business statistics infrastructure. A subsequent update in 2005 further streamlined processes and strengthened the foundation for the UES framework. Further enhancements to the business statistics infrastructure were introduced in 2010 with the launch of the Integrated Business Statistics Program (IBSP), which aimed to increase the efficiency, consistency, and responsiveness of business surveys (Daoust, Mireuta and Kleim, 2023). The IBSP was designed with six core objectives: i. to improve data quality; ii. to reduce response burden through coordinated sampling strategies; iii. to modernize data processing infrastructure; iv. to integrate the majority of economic surveys under a unified framework; v. to standardize processes and tools across programs; and vi. to enhance adaptability to evolving data needs.

A similar modernization effort was explored by the U.S. Census Bureau in 2018, as documented by Kirkendall, White Jr, Citro and Abraham (2018), reflecting a broader international trend toward more integrated and efficient economic data collection systems.

2.2 Structure of the BR

The BR is a key resource for the conduct of business surveys and the production of reliable economic statistics in Canada. Its primary function is to serve as a high-quality sampling frame for the economic surveys administered by Statistics Canada. The BR encompasses all economic entities engaged in the production of goods and services, including incorporated and unincorporated enterprises, commercial firms, non-profit organizations, religious institutions, and government bodies.

In statistical terms, a business is defined as an economic unit, such as an establishment or farm, that utilizes resources (labor, capital, and raw materials) to produce goods or services. Businesses operate across a diverse range of economic sectors, including but not limited to retail trade, wholesale trade, services, manufacturing, construction, energy, transportation, agriculture, and international trade. A business survey collects data used for statistical purposes from a sample of businesses or firms.

Each business in the BR is described by a series of attributes grouped into five main categories: identification data, classification data, contact data, activity status, and maintenance/linkage data.

- Identification data uniquely characterize each unit via elements such as business name, physical address, and alphanumeric identifiers (the BN).
- Classification data include size, industrial, and geographic categorizations, which are essential for stratifying the population and drawing representative samples.

- Contact data has details necessary for survey administration, such as contact persons, mailing addresses, telephone numbers, and past survey response history.
- Activity status denotes whether a unit is currently operational (i.e., active, live, in-season) or inactive (i.e., dead, out-of-season).
- Maintenance and linkage data are used to track the temporal evolution of business entities, capturing changes in structure or activity through linkage records and change histories.

Collectively, these components constitute the frame data required to support business survey operations. A business is included in the BR if it meets at least one of the following criteria:

1. It remits payroll deductions to the CRA for its employees.
2. It reports at least $30,000 in annual gross revenue.
3. It is incorporated under federal or provincial legislation and has filed a federal corporate income tax return (T2) within the past three years.

The BR is constructed and maintained using data from three principal administrative sources:

- Payroll Deductions (PD): Monthly or quarterly employer remittances to CRA for employment insurance, pension contributions, and other payroll-related obligations.
- Goods and Services Tax (GST) accounts: Records of GST remittances and input tax credit claims by registered entities.
- Income tax returns: Annual tax filings from incorporated businesses (T2) and unincorporated businesses (T1), detailing revenue, expenditures, and other financial information.

The integration of these administrative sources is facilitated by the BN, as described in Section 2.1. The BN serves as a unique identifier that links PD, GST, and income tax data, thus minimizing duplication and supporting the consistent representation of legal and operational structures within the BR.

The BR distinguishes between *legal structures* and *operational structures* to more accurately reflect the complex ways in which businesses function. The *legal structure* defines the entity's legal identity, which governs interactions with government institutions (e.g., tax filings and payroll administration) and shapes contracts, ownership, and labor relations. The *operational structure* reflects the internal organization of the business, encompassing how it allocates resources, delivers goods and services, and maintains its accounting practices.

For small businesses, legal and operational structures are typically aligned; however, larger enterprises often operate through multiple legal entities across various lines of business. The BR accounts for this complexity by linking multiple units into a consolidated structure, based on the type of information each unit can provide. This allows for flexible sampling and data collection strategies tailored to the nature of the business entity and the survey's objectives.

2.3 Statistical units

The BR standardizes business structures into well-defined *statistical units*, which serve as the foundational elements for economic measurement and survey design. This transformation is guided by key principles such as decision-making autonomy, homogeneity of industrial activity, and the availability of data. The delineation of these units allows for consistent statistical representation of the business population across various surveys and analytical frameworks.

The BR uses a four-level hierarchy of statistical units. Each level reflects a different aspect of the organizational and operational structure of businesses and supports specific data collection and analysis needs:

- *Enterprise*: This is the highest statistical unit within the operational structure and represents the complete legal entity. The enterprise corresponds to a unit capable of producing a comprehensive set of financial statements, encompassing the entirety of its economic activities. It serves as the primary unit for collecting aggregate financial information, including assets, liabilities, revenues, and expenditures.
- *Company*: This unit corresponds to the lowest level at which operating profit can be measured. It also provides detailed information on assets and liabilities, enabling the measurement of capital employed. The company unit is particularly relevant for financial analysis and performance assessment at a managerial or divisional level.
- *Establishment*: This unit refers to the smallest grouping of production entities that maintains a degree of organizational autonomy and typically operates from a single geographic location. Establishments are characterized by the production of a relatively homogeneous set of goods or services. They do not span provincial or territorial boundaries and are capable of reporting data on output values, principal input costs, and labor expenditures. The establishment is often the preferred unit for industry-specific surveys and production analysis.
- *Location*: The location unit is a subdivision of the establishment, representing a single physical site or a defined group of closely situated sites (e.g., a plant, warehouse, or retail outlet). A location can report selected operational variables such as revenue and, in many cases, employment. This level of granularity is particularly useful for regional or localized economic analyses.

For complex businesses, where multiple activities or geographic regions are involved, each statistical unit is assigned a dominant industry classification code to ensure appropriate categorization. Enterprises are classified based on the primary industry associated with their overall output, while establishments receive both industry and geographic classification codes reflecting their principal economic activity, typically determined by revenue. These classifications are grounded in the North American Industry Classification System (NAICS) and are regularly updated to maintain relevance and accuracy.

The BR prioritizes the maintenance of classification accuracy for larger and more complex business entities, as errors at these levels can have significant implications for statistical outputs. Continuous

monitoring and refinement of industry and geographic classifications ensure the integrity and representativeness of the sampling frame used for business surveys.

2.4 Data collection arrangements

Data collection arrangements between the statistical agency and a sampled business (defined at the statistical unit level) are established via collection units. Three attributes associated with a collection unit are:

- *Coverage*: defines the relationship between the business from which the data are being acquired and the level within the business (i.e., enterprise, location) for which the data are required;
- *Collection mode*: the means of obtaining the data (e.g., questionnaire, telephone interview, administrative record, etc.);
- *Contact information for the respondent*: name, physical or email address, and telephone number within the business operating structure.

Figures 2.1 and 2.2 illustrate how statistical units and collection units are related to a simple or complex business, respectively. Collection units provide one of several means for updating the business frame in terms of frame data: others include administrative data updates and profiling. Collection units also represent the vehicle for monitoring respondent contacts and assessing respondent burden. Collection units are created only for statistical units in a survey sample and are survey-specific. Collection units are automatically generated with rules that depend on statistical-operating links and data availability. They can be modified manually, as needed, to take into account information related to nonstandard reporting arrangements requested by respondents.

Figure 2.1 Simple business

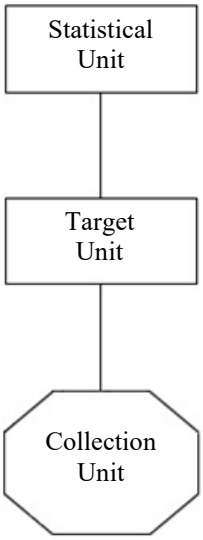
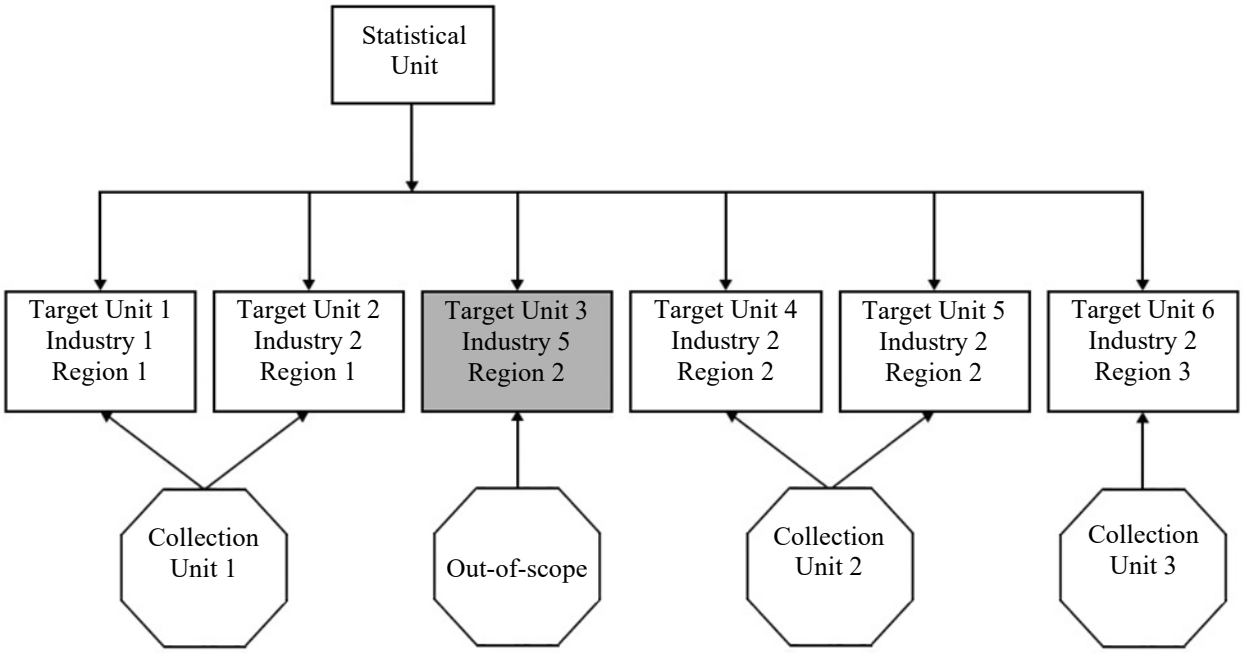


Figure 2.2 Complex business



2.5 Maintenance

Businesses are constantly evolving, opening, closing, merging, and changing in terms of size, industry, and location. These shifts can affect classification information, either in the sampled or non-sampled portions of the population. Such changes may reflect real changes in the business's activities, or they could be due to prior misclassification. Changes in industry, location, or size are more commonly detected in businesses already part of a sample (in-sample) compared to those not in a sample (out-of-sample). In one-time surveys, if a classification change occurs after the sample has been selected, the original weight of the business is retained for estimation purposes, but the updated classification is used for reporting. For repeated surveys, businesses may need to be reclassified over time, and domain estimation can help adjust for these changes on multiple occasions.

The BR regularly updates the industry and geographic classifications of all its business entities, focusing particularly on larger and more complex businesses. The register also maintains information on the production and accounting entities of these businesses.

- *Births*: New units enter the BR primarily through PD, GST, or corporation accounts. CRA assigns a BN to these units and requests legal and operational details. For established businesses acquiring additional accounts, new accounts are linked to their BN root.
- *Deaths*: Inactive businesses are deactivated from the BR based on remittance activity. For simple businesses, deactivation occurs if the business stops remitting to the CRA for 12 months. For complex businesses, deactivation happens only when all associated accounts cease to remit. If some accounts stop remitting, the BR triggers a reprofiling to update the business's structure and contact information.

- *Size Updates*: Size measures, such as the number of employees and gross business income, are updated using models based on remittances. For simple businesses, PD and GST data help update size measures. For complex businesses, economic surveys are used for size updates.
- *Surveys*: Economic surveys using the BR frame signal changes in business structures, contact information, and size measures. The BR surveys any unit with discrepancies between frame data and collected survey data. It also surveys complex businesses in response to signals from administrative sources or as part of periodic updates, such as the Profiling Survey. Profiling is the process of conducting in-depth telephone or on-site interviews with senior company representatives to obtain all pertinent financial information, relationships, and structures about the company (see College, 1995). The collected data updates the BR's operational, legal, and accounting entities, as well as its statistical structures.

3. Sample size determination and allocation

As outlined in Section 2, the BR organizes the business population into a hierarchy of four statistical units: enterprise, company, establishment, and location. Depending on the objectives and analytical requirements of a given survey, the target universe may be defined at any of these levels. Once the target universe has been delineated, a corresponding set of target units is identified to form the basis for sample design.

Sampling within the business universe typically involves two main steps. The first step is to define the in-scope target universe and identify the relevant statistical units. The second step is the selection of the sampling unit, which must be at least as detailed as the target unit. For example, if both locations and establishments are defined as statistical units in the BR and the target unit is the location, the sampling unit may be either the location itself or the more aggregated establishment level. The choice between them depends on operational and methodological considerations. The selection of the sampling unit carries important implications for data collection, estimation procedures, and overall survey logistics. It influences not only the level of detail in the collected data but also the practicality of linking survey responses with administrative or auxiliary data sources.

3.1 Stratification of sampling units

After the sampling unit has been identified, the business population is stratified to enhance the efficiency and precision of the sampling design. This stratification is typically implemented in two stages:

- Primary stratification is based on key structural variables, most commonly *geographic region* (e.g., provinces, territories, and major metropolitan areas) and *industry classification* (e.g., retail trade, accommodation and food services, finance and insurance, or manufacturing), often based on the NAICS code.
- Secondary stratification is based on measures of business size, which may include variables such as the number of employees, annual gross revenue, or total sales. This stratification reflects the

heterogeneity of business units and the right-skewed distribution of economic activity, where a small number of large businesses account for a disproportionate share of the total economic output.

Given the pronounced skewness of business populations, stratification often involves the creation of a "take-all" stratum, in which all units are selected with certainty due to their size or importance. The remaining units are divided into one or more "take-some" strata, from which samples are drawn probabilistically. Two principal approaches are considered:

- The n-scenario, which begins with a fixed total sample size of n, is typically constrained by budgetary or operational limitations. Under this scenario, the goal is to allocate the available sample optimally across strata to maximize precision within the given resources.
- The c-scenario, which begins with a specified target level of precision, is expressed as an acceptable coefficient of variation (CV) for key estimates. In this case, the required sample size is calculated to achieve the desired CV. If the resulting sample size exceeds available resources, the CV target must be relaxed accordingly to align with budget constraints.

Once either the n-scenario or c-scenario is selected, sample sizes for each secondary stratum can be determined using allocation methods such as Neyman allocation, power allocation, or other optimal techniques designed to minimize variance or maximize efficiency, while accounting for cost structures and response burden.

3.1.1 The *n-scenario*: Allocating the overall sample size to the primary and secondary strata

Sample sizes are allocated to the primary strata using the univariate allocation method proposed by Bankier (1988). For a given stratification of the universe U let the strata be denoted as U_{gi}, where $g = 1,\ldots,G$ and $i = 1,\ldots,I$. Let their respective population sizes be N_{gi}. Their union is represented as $U = \bigcup_{g=1}^{G} \bigcup_{i=1}^{I} U_{gi}$. Stratified simple random samples s_{gi} of size n_{gi} are selected from U_{gi} where $g = 1,\ldots,G$ and $i = 1,\ldots,I$. The stratum sample sizes n_{gi} minimize the loss function.

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

subject to the constraint $n = \sum_{g=1}^{G} \sum_{i=1}^{I} n_{gi}$.

In equation (3.1), $\mathrm{CV}(\hat{Y}_{gi}) = V(\hat{Y}_{gi}) \big/ Y_{gi}^2$, and X_{gi}^{q} is a measure of the size or importance of the gith primary stratum: q is a constant in the range $0 \le q \le 1$. The population total of the gith primary stratum is $Y_{gi} = \sum_{\ell \in U_{gi}} y_{gi\ell}$, and $y_{gi\ell}$ is the ℓth value of variable y in stratum gi. The estimator of Y_{gi} is $\hat{Y}_{gi} = N_{gi} \sum_{\ell \in s_{gi}} y_{gi\ell} \big/ n_{gi}$. It can be shown that the function F_1 is minimized by

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

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

When $q = 1$ and letting $X_h = Y_h$, we obtain the Neyman allocation

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

On the other hand, when $q = 0$ and $S_{gi} / \overline{Y}_{gi}$ do not vary significantly from stratum to stratum and the FPC factors can be ignored, we obtain an equal allocation between the strata: that is $n_{gi} = n / GI$. In this case, the CVs for the different strata are almost equal. As Bankier (1988) pointed out, Neyman allocation minimizes the coefficient of variation (CV) across the primary strata. However, some primary stratum may have estimators of the totals that have large CVs. Alternatively, an allocation that achieves nearly equal CVs for the primary stratum estimators may result in overall estimators having a much larger CV than under Neyman allocation.

Bankier (1988) recommended $q = 0.3$ or 0.5 as a compromise between these two types of allocations. Other vues of q can be used, provided that they are in the interval [0, 1]. Note that n_{gi} is fixed for each combination of g and i. A multivariate extension to Bankier (1988) is given in Appendix A.

Assuming that there is an auxiliary variable x that is highly correlated with y for each population element, we can stratify each primary stratum, U_{gi}, into secondary-size strata. Without loss of generality, we consider the case where we have a single primary stratum. The extension to multiple primary strata is straightforward since strata are independent. To this end, let a given primary stratum be denoted as U and a sample selected from it as s. The corresponding population and sample sizes are denoted as N and n, respectively. Each primary stratum U is then stratified into H-size strata $U_1, U_2, \ldots, U_H$ with corresponding population counts $N_1, N_2, \ldots, N_H$. The samples selected from these strata are denoted as s_1, $s_2, \ldots, s_H$, and the corresponding sample sizes as $n_1, n_2, \ldots, n_H$. The associated stratification boundaries are denoted as $b_0, b_1, \ldots, b_{H-1}, b_H$, where $b_0 < b_1, \ldots < b_{H-1} < b_H$: b_0 and b_H represent the minimum and maximum values of x, respectively. If no "take-all" units are specified, the estimator of the population mean is

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

For a given primary stratum, the size boundaries are obtained by minimizing

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

where $n_h = n a_h$, $A_h = N_h^2 S_h^2$, $D = \sum_{h=1}^{H} N_h S_h^2$ and S_h^2 is the variance of the known x values within the hth secondary stratum within a given primary stratum gi. The term a_h is a general-purpose allocation formula given by Hidiroglou and Srinath (1993). It is given by

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

where $0 \le 2q_1 \le 1$, $0 \le 2q_2 \le 1$, and $0 \le 2q_3 \le 1$. This general formula covers all known allocation schemes, including the power allocation scheme given by Bankier (1988). Table 3.1 provides a_h for Neyman, X-proportional, N-proportional, and the power allocation given by Bankier (1988).

Table 3.1
Correspondence between $a_h, h = 1,2,\ldots,H,$ and known allocation schemes

Allocations Scheme	$a_h, h = 1,2,\ldots,H,$	$2q_1$	$2q_2$	$2q_3$
Neyman	$N_h S_h \Big/ \left(\sum_{h=1}^{H} N_h S_h\right)$	1	0	1
X -proportional	$X_h \Big/ \left(\sum_{h=1}^{H} X_h\right)$	0	1	0
N -proportional	$N_h \Big/ \sum_{\ell=1}^{L} N_h$	1	0	0
Power allocation	$X_h^q \Big/ \sum_{h=1}^{H} X_h^q;\ 0 \le q \le 1$	q	q	0

Using Dalenius' (1950) approach, $V(\hat{\bar{X}}_1)$ is transformed from an expression based on a finite population to one applicable to a continuous population. Specifically, given a continuous density function $g(x)$ of the auxiliary variable x in the range $(-\infty,\infty)$, the conditional mean and variance of the hth stratum U_h, $h = 1,\ldots,H$ can be expressed as $\mu_h = \int_{b_{h-1}}^{b_h} xg(x)\,dx/\omega_h$, and $\sigma_h^2 = \int_{b_{h-1}}^{b_h} x^2 g(x)\,dx/\omega_h - \mu_h^2$ where $\omega_h = \int_{b_{h-1}}^{b_h} g(x)\,dx$. The continuous version of $V(\hat{\bar{X}}_1)$ is derived as follows:

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

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

The optimum boundaries, $b_1, b_2, \ldots, b_{H-1}, b_H$, where $b_2 \le \ldots \le b_{H-1} \le b_H \le x_{(N)}$, are determined by taking the partial derivatives of (3.6) for each $b_h, h = 1,\ldots,H,$ setting them to zero, and solving the resulting quadratic equations iteratively using a procedure such as the one suggested by Sethi (1963). The initial boundary values are set by assigning an equal number of elements to each group. Further details on the derivation process can be found in Lavallée and Hidiroglou (1988), Baillargeon, Rivest and Ferland (2007), and Hidiroglou and Kozak (2018). The R program developed by Rivest and Baillargeon (2022) performs the necessary computations to obtain the boundaries for this case.

If we choose to include a take-all stratum, it will be denoted as U_H, and the associated take-some strata as $U_1, U_2, \ldots, U_{H-1}$. The estimator of the mean $\bar{X}$ is

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

The population variance of $\hat{\bar{X}}_2$ is

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

The continuous analog of (3.7) in terms of the continuous distribution $g(x)$ is

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

where $A, B,$ and D have been previously defined. The stratum boundaries are obtained by differentiating (3.8) with respect to b_h, $h = 1, 2, \ldots, H-1,$ and setting the resulting equation to zero.

3.1.2 The c -scenario: Allocating the overall expected coefficient of variation to the primary strata

The required sample size for each primary stratum is determined using an expected coefficient of variation (c) at the national level. For a given primary stratum $U_{gi}, g = 1, \ldots, G,\ i = 1, \ldots, I,$ let $\hat{X}_{gi} = N_{gi} \sum_{k \in s_{gi}} x_k / n_{gi}$ be the estimator of the population total $X_{gi} = \sum_{k \in U_{gi}} x_k,$ where

$N_{gi} =$ number of population units,

$n_{gi} =$ number of units to be sampled,

and

$S_{gi}^2 =$ population variance for x.

Let the variance of the overall estimated total for $x,$ $\hat{X}_{..} = \sum_{i=1}^{G} \sum_{i=1}^{I} \hat{X}_{gi},$ be given by $V(\hat{X}_{..})$. This variance can also be expressed as $V(\hat{X}_{..}) = c^2 X_{..}^2,$ where $X_{..} = \sum_{g=1}^{G} \sum_{i=1}^{I} X_{gi}$ and c is the expected overall coefficient of variation.

The national coefficient of variation c is allocated to the primary strata (province by industry) using a raking procedure (see Deming and Stephan, 1940). The steps are as follows. The marginal coefficients of variation for each geographical level $c_{g.},$ and industrial level $c_{.i},$ are obtained by assuming that:

i. the $c_{g.}$'s are all equal for $g = 1, \ldots, G,$ and
ii. the $c_{.i}$'s are all equal for $i = 1, \ldots, I$.

The $c_{g.}$'s and $c_{.i}$'s obey the following two conditions:

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

and

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

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

We first obtain the marginal coefficients of variations $c_{g.}$ (coefficient of variation for the gth geographic stratum $U_{g.}$) and $c_{.i}$ (coefficient of variation for the ith industrial stratum $U_{.i}$). Since $V(\hat{X}_{g.}) = \left(c_{g.} X_{g.}\right)^2,$ where $X_{g.} = \sum_{i=1}^{I} X_{gi},$ using condition iii. results in $\sum_{g=1}^{G} (c_{g.} X_{g.})^2 = c^2 X_{..}^2$. Using condition i., we obtain that

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

Similarly, using conditions ii. and iv., we obtain

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

An initial coefficient of variation for the gith primary stratum is set to $c_{gi}^{(0)} = 0.5\ (c_{g.} + c_{.i})$. The next step is to iterate the coefficients of variation as follows:

$$c_{gi}^{(r)} = c_{gi}^{(r-1)} \begin{cases} \dfrac{\left(c_{.i} X_{.i}\right)}{\sqrt{\sum_{i=1}^{I} c_{gi}^{(r-1)} X_{gi}^2}} & \text{if } r \text{ is odd} \\ \dfrac{\left(c_{g.} X_{g.}\right)}{\sqrt{\sum_{g=1}^{G} c_{gi}^{(r-1)} X_{gi}^2}} & \text{if } r \text{ is even} \end{cases}$$

for $r = 1,\ldots,R,$ until the following convergence criterion is met: $\left| c_{gi}^{(t)} - c_{gi}^{(t+1)} \right| < \varepsilon$ for some $t \geq 5$ and a small ε.

In practice, five iterations are sufficient to stabilize the $c_{gi}^{(r)}$ values. We denote the final coefficients of variation as $c_{gi}^{(f)}$. We then proceed as in Section 3.1.1 to stratify the individual primary strata $gi,$ using the same notation.

If all the size strata are take-some, the estimator of $\bar{X} = N^{-1} \sum_{h=1}^{H} \sum_{k \in U_h} x_k$ is given by equation (3.4). For a given primary stratum, the size boundaries for this case are obtained by minimizing

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

where $W_h = N_h / N$ and $c = c_{gi}^{(f)}$ is the required coefficient of variation for primary stratum $gi,$ and the other terms of equation (3.9) have been defined in Section 3.1.1. The continuous analogue of (3.9) is

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

where

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

If we have a take-all stratum and the remaining strata are take-some, the estimator of $\bar{X} = \sum_{h=1}^{H} \sum_{k \in U_h} x_k / N$ is given by equation (3.6). The corresponding sample size is

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

The continuous analogue of (3.10) is

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

where

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

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

For a given reliability c and allocation $a_h (h = 1,\ldots,H-1),$ let the resulting boundaries be $b_1^{(c)}$, $b_2^{(c)},\ldots,b_{H-1}^{(c)},$ and the associated minimum overall sample size be $n.$ Using this sample size n and the same allocation a_h as input to the n scenario, we obtain boundaries $b_1^{(n)}, b_2^{(n)},\ldots,b_{H-1}^{(n)}$ and an expected reliability $c^{(n)}$. The c and n scenarios are equivalent, that is $b_h^{(n)} = b_h^{(c)},\ h = 1,\ldots,H-1,$ and $c^{(n)} = c.$ The proof is given in Appendix B.

3.2 Extensions to the optimal boundary problem

The impact of an out-of-date frame should be reflected in the sample size determination and allocation method. The out-of-date nature of a frame occurs because the classification (geography, industry, size variable, or live or dead status) of the units is not up to date. In our case, we just focus on estimating the total of the live units for a variable $y,$ given that dead units are present on the frame but identified as active. Consequently, a representative portion of them will be included in the sample. The universe of "active" units is labeled as $F_A.$ The corresponding universe of live units (but unknown) is denoted as F_L $F_L \subset F_A$. Let the parameter of interest be the domain total $X_L = \sum_{k \in F_A} x_{k\ell},$ where $x_{k\ell}$ is equal to x_k if $k \in F_L$ and zero otherwise. The estimator of the total X_L is $\hat{X}_L = \sum_{h=1}^{H} \hat{X}_{hL}$ with $\hat{X}_{hL} = \sum_{k \in s_{Ah}} (N_{Ah} / n_{Ah})\, x_{k\ell}.$ We need to determine the sample size $n_A,$ such that: (i) the allocation to the design strata is $n_{Ah} = n_A\, a_h (0 < a_h < 1),$ and (ii) the targeted coefficient of variation c is satisfied, that is, $V(\hat{X}_L) = c^2 X_L^2.$ The population F_A is stratified into H strata $F_{A,h},$ each of size $N_{Ah}.$ Simple random samples s_{Ah} of size n_{Ah} are selected from each stratum $F_{A,h}, h = 1,\ldots,H,$ without replacement. Let $F_{L,h}, h = 1,\ldots,H$ denote the corresponding set of live units within stratum $F_{A,h}.$ Define

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

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

Using $V(\hat{X}_L)$ the required overall sample size n_A is

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

where $\tilde{S}_{Lh}^2 = (N_{Lh} - 1)^{-1} \sum_{k \in F_{hL}} (x_k - \tilde{\bar{Y}}_{Lh})^2, \tilde{\bar{Y}}_{Lh} = \sum_{k \in F_{Ah}} x_{k\ell} / N_{Lh},$ N_{Lh} is the number of units belonging to the domain $F_{Lh} = F_L \cap F_{Ah},$ and $P_{Lh} = N_{Lh} / N_{Ah}$ is the expected proportion of units that belong to U_ℓ and that were initially sampled in stratum $h.$

Note that when $P_{h\ell} = 1$ we obtain equation (3.9). The mean and variance components can be estimated from previous surveys. The required sample sizes at the stratum level are then simply $n_{Ah} = n_A\, a_h$ for

$h = 1, \ldots, H$. It is not recommended to use an approximation of the type $n^*_{Ah} = n_{Ah}\, a_h$ to compensate for unknown dead units, where n_h is computed ignoring the existence of unknown dead units in the universe.

Rivest (2002) expanded on the Lavallée-Hidiroglou (1988) stratification algorithm by addressing the distinction between the stratification and survey variables. He introduced a log-linear model-based stratification algorithm, which he applied to several surveys developed by the Statistical Consulting Unit at Université Laval. Baillargeon and Rivest (2009) built on this work by incorporating the option of a "take-none" stratum and accounting for uniform non-response probabilities across size strata. Additionally, they demonstrated that, from a computational standpoint, the algorithm proposed by Kozak (2004) was the most efficient tool for solving the optimization problem associated with determining stratum boundaries. Further contributions from Hidiroglou and Laniel (2001) and Lednicki and Wieczorkowski (2003) suggested incorporating the anticipated moments of the variable of interest y, conditional on an auxiliary variable x, into the stratification process. There should be a minimum number of units within each take-some stratum to ensure that non-response does not result in empty strata. A maximum weight should be set to avoid the negative impact of outliers. The application of the above additional rules is well illustrated in Ferland and Batten (2007).

3.3 Numerical comparison between the allocations to the primary strata

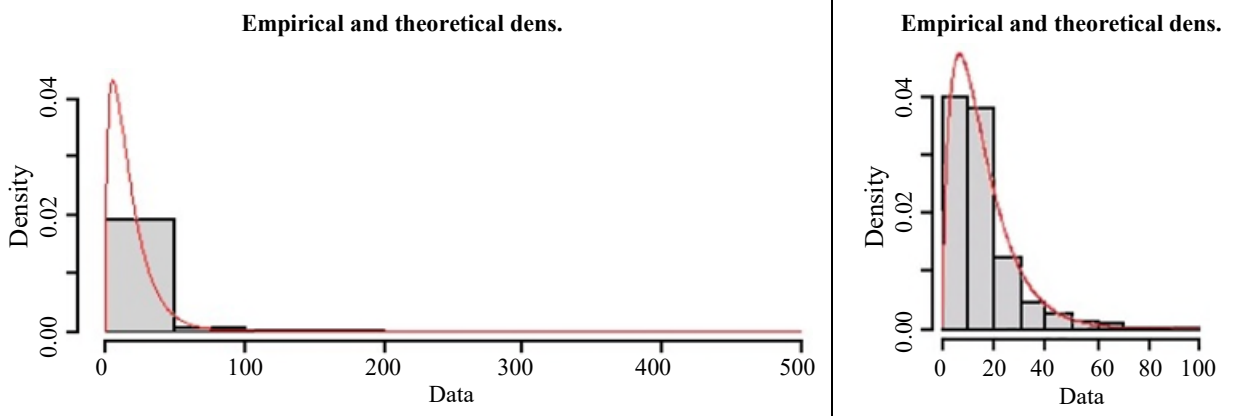
The methods given in Sections 3.1.1 and 3.1.2 are illustrated by creating data sets based on the one given in Rivest and Baillargeon (2022). This data set was created using a "simulated" size measure for the Monthly Retail Trade Survey (MRTS) conducted by Statistics Canada. This dataset had 2,000 establishments. The size measure for each of these establishments was created by combining the MRTS data with three administrative variables from corporation tax returns. Due to the highly skewed nature of the data, the twenty-one largest observations were deleted to obtain a converging fit of the gamma distribution. The gamma distribution is given by $f(x) = (\Gamma(\alpha))^{-1} \beta^{\alpha} x^{\alpha-1} e^{-\beta x}$, where α and β are its respective shape and rate parameters. These parameters were estimated as $\hat{\alpha} =$ 1.7610(0.0515) and $\hat{\beta} =$ 0.1147(0.0039): the values in the brackets are their corresponding standard errors. Figure 3.1 provides summary statistics for both the original and truncated data.

To evaluate the impact of different allocation methods on the stratified sample design, a numerical comparison was conducted for fifteen primary strata. These strata were defined through a cross-classification of five Canadian provinces and three industry groups. The choice of these particular industries and provinces was motivated by their economic significance and sectoral diversity, which enables meaningful comparison across strata. The provinces included Quebec, Ontario, Manitoba, Alberta, and British Columbia. The industry groups were based on the three-digit NAICS level. The selected NAICS were

- 441: Motor Vehicle and Parts Dealers,
- 444: Building Material and Garden Equipment and Supplies Dealers, and
- 445: Food and Beverage Stores.

Figure 3.1 Summary statistics

Original MRTS data	Truncated MRTS data
n = 2,000: skewness 8.622	n = 1,979: skewness 2.315



Note: MRTS = Monthly Retail Trade Survey.

Tables 3.2 and 3.3 present the number of establishments and their corresponding total sales, respectively, for each of the fifteen primary strata.

Table 3.2
Number of retail establishments (N_{gi}) for each primary stratum, $g = 1,\ldots,5$ and $i = 1, 2, 3$

Province	Industry			
	NAICS 441	NAICS 444	NAICS 445	Across industries
Quebec	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
British Columbia	1,826	1,220	3,707	6,753
Across provinces	10,795	7,037	25,341	43,173

Note: NAICS = North American Industry Classification System.
Source: Extracted from Statistics Canada. Table 33-10-0806-01 Canadian Business Counts, with employees, June 2024 https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3310076101.

Table 3.3
Total Sales (X_{gi}) in thousands of dollars (unadjusted for seasonality) for each primary stratum, $g = 1,\ldots,5$ and $i = 1, 2, 3$

Province	Industry			
	NAICS 441	NAICS 444	NAICS 445	Across industries
Quebec	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
British Columbia	1,969,426	567,892	1,992,689	4,530,007
Across provinces	16,806,510	4,007,749	12,035,916	32,850,175

Note: NAICS = North American Industry Classification System.
Source: Extracted from Statistics Canada. Table 20-10-0056-01 Monthly retail trade sales by province and territory (× 1,000), June 2024.

Next, the coefficients of variation for the provincial by industry totals were computed. These were obtained using the raking procedure given in Section 3.1.2. The overall coefficient of variation, c, was set to 1%. The resulting coefficients of variation are displayed in Table 3.4.

Table 3.4
Raked coefficients of variation $c_{gi}^{(f)}$ (%), by province and industry, $g = 1,\ldots,5$ and $i = 1, 2, 3$

Province	Industry			
	NAICS 441	NAICS 444	NAICS 445	Across industries
Quebec	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%
British Columbia	2.91%	2.94%	2.94%	1.89%
Across provinces	1.54%	1.54%	1.54%	1.00%

Note: NAICS = North American Industry Classification System.

The individual observations for each of these primary strata were generated using the following inputs.

1. The number of establishments (as reported in Table 3.2),
2. Their total sales (as reported in Table 3.3), and
3. The estimated parameters of the gamma distribution where $\hat{\alpha} = 1.7610$ and $\hat{\beta} = 0.1147$.

The generated values were then prorated to ensure that their aggregates matched the total sales figures reported in Table 3.3. Using these generated values, the primary strata were stratified using three sample allocation methods. These were:

1. Two take-some strata, using Neyman allocation.
2. A specified take-all stratum with a single take-some stratum, and
3. A specified take-all stratum with two take-some strata using Neyman allocation.

The third method was selected, as it resulted in the smallest number of units in every case.

Using the coefficients of variation in Table 3.4 for each primary stratum, the required sample sizes are presented in Table 3.5.

Table 3.5
Sample sizes (n_{gi}) by province and industry with specified take-all and two take-some strata (Neyman allocation), $g = 1,\ldots,5$ and $i = 1, 2, 3$

Province	Industry			
	NAICS 441	NAICS 444	NAICS 445	Across industries
Quebec	183	170	194	547
Ontario	189	175	201	565
Manitoba	139	90	134	363
Alberta	167	130	191	488
British Columbia	157	147	185	489
Across provinces	835	712	905	2,452

Note: NAICS = North American Industry Classification System.

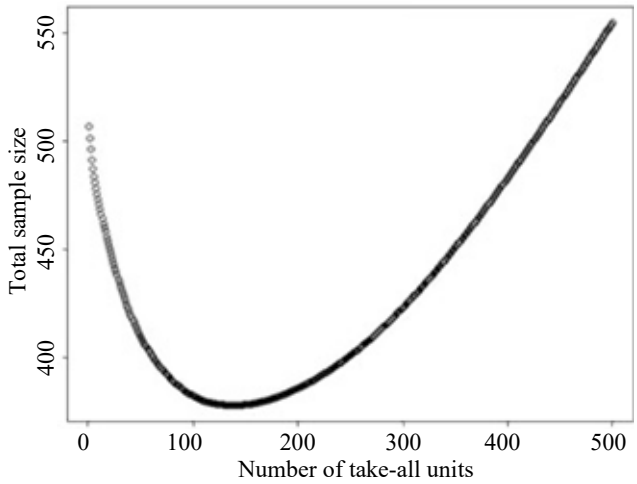
Note that the specification of a take-all stratum and a single take-some stratum may lead to a larger overall sample size compared to using two take-some strata with Neyman allocation. For instance, the sample size for Motor Vehicle and Parts Dealers in Alberta would have been 191 with two take-some strata

using Neyman allocation, whereas it would be 378 units with a combination of a take-all and a take-some stratum. This result, illustrated in Figure 3.2, was confirmed using the exact formula provided by Hidiroglou (1986b), given by

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

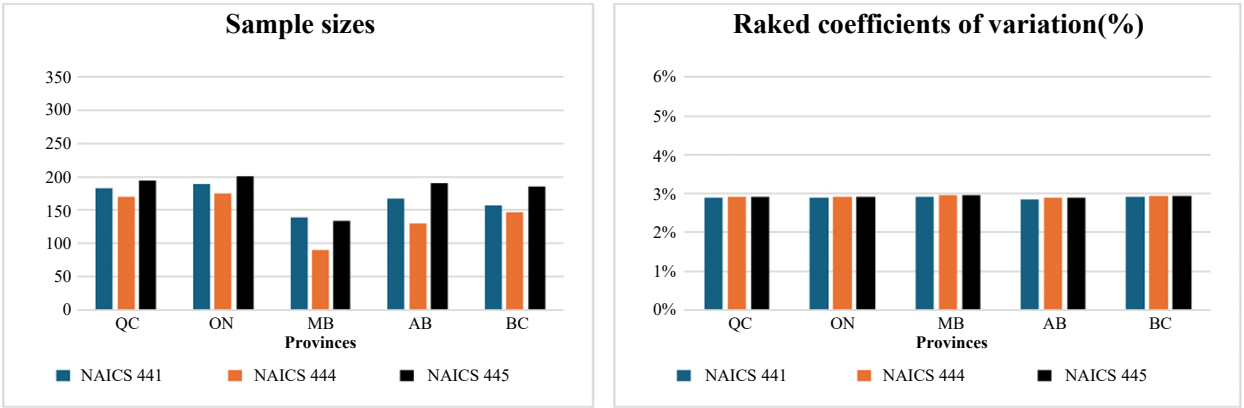
where $n(t)$ is the total number of units to be sampled, including the take-all units; t is the number of take-all units; $S^2_{[N-t]}$ is the population variance of the smallest $N-t$ units, and c is the specified coefficient of variation.

Figure 3.2 Sample size versus number of take-all units



A graphical representation of the results in Tables 3.4 and 3.5 is given in Figure 3.3.

Figure 3.3 Primary strata: sample sizes (n_{gi}), and corresponding raked coefficients of variation ($c_{gi}^{(f)}$), $g = 1,\ldots,5$ and $i = 1, 2, 3$



Note: NAICS = North American Industry Classification System.

Next, the sample sizes $n_{gi} = nX_{gi}^{q} \Big/ \sum_{g=1}^{G}\sum_{i=1}^{I} X_{gi}^{q}$, $g = 1,\ldots,5$ and $i = 1, 2, 3$ where $0 \le q \le 1$, were computed using Bankier's power allocation using equation (3.3). The overall sample size for the fifteen primary strata was $n = 2{,}452$ (see Table 3.5). This allows for a comparison of the resulting coefficients of variation and sampling fractions for each of the fifteen primary strata with those obtained using the raking procedure summarized in Figure 3.3.

The two extremes for q are 0 and 1. When $q = 0$, we have an equal allocation of the sample to each of the primary strata: that is $n_{gi} = n / N$ $g = 1,\ldots,G$ and $i = 1,\ldots,I$. When $q = 1$, we have Neyman allocation of the sample to each of the primary strata: that is $n_{gi} = n / N$ $g = 1,\ldots,G$ and $i = 1,\ldots,I$: that is

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

As we will see later, neither of these choices is ideal. The stratification of the 15 primary strata into secondary strata used the same size stratification as the one given in Table 3.5: that is, a take-all stratum and two take-some strata (using Neyman allocation).

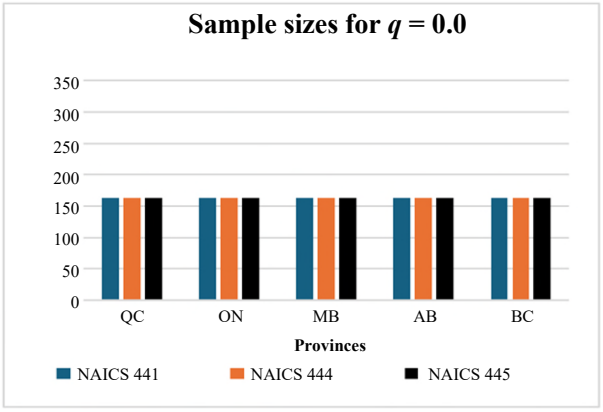
We compared the sampling fractions and coefficients of variation (CVs) using the raking procedure with five different values for the selection of the q values (0.0, 0.1, 0.3, 0.5, and 1.0) in Bankier's (1988) procedure. These comparisons were based on a total sample size of 2,452, calculated using the raking procedure outlined in Section 3.2.2. A graphical summary of the results is provided in Figure 3.4 for the various q's.

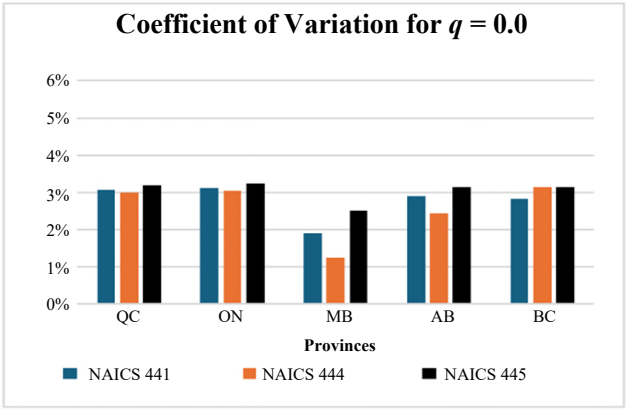
Several observations can be drawn from Figure 3.4.

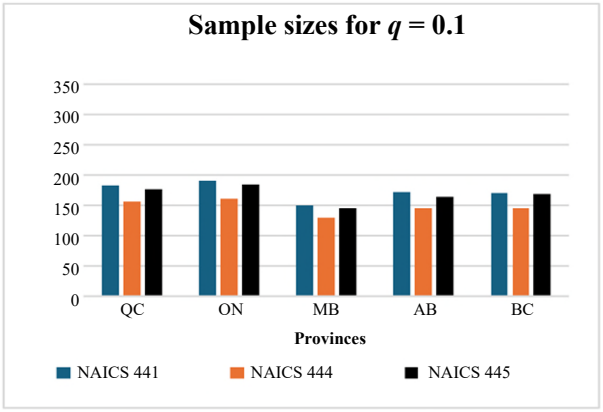
- *Sample sizes*: The sample sizes increase as q increases in the provinces with the largest sales, Quebec and Ontario. The sample size increase is moderate in the next two provinces with the largest sales as q increases, Alberta and British Columbia. This is true for NAICS 441 and 445. However, they decrease for NAICS 444 as q increases in all provinces. The sample size decreases as q increases for the smallest province, Manitoba, for all the NAICS. This holds for both the raking method (Figure 3.3) and Bankier's (1988) allocation (Figure 3.4). When $q = 0$, the sample size is the same for all primary strata, that is, 163. When $q = 1$, the sample size increases as the size of the primary strata increases.
- *Coefficients of variation*: The coefficients of variation decrease as q increases in the provinces with the largest sales, Quebec and Ontario. The coefficients of variation decrease moderately as q increases in the next two provinces with the largest sales, Alberta and British Columbia. This is true for NAICS 441 and 445. However, they increase for NAICS 444 as q increases in all provinces. The coefficients of variation decrease as q increases for the smallest province, Manitoba for all the NAICS. This holds for both the raking method (Figure 3.3) and Bankier's (1988) allocation (Figure 3.4). When $q = 0$, the coefficients of variation vary between 1.25% to 3.12%. When $q = 1$, the coefficients of variation decrease as the size of the primary strata increases.

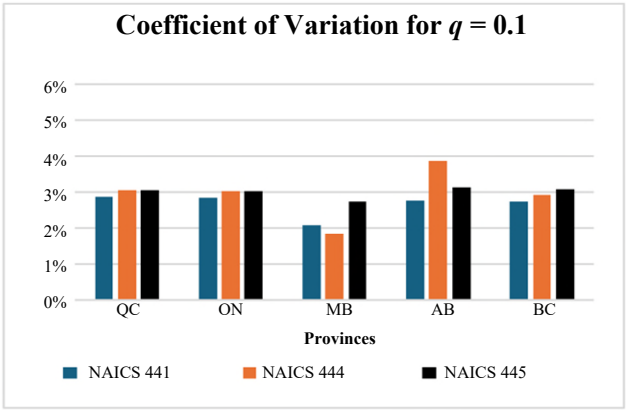
The values for q in the range 0.1 to 0.3 offer a reasonable trade-off between sample sizes and CVs. Although the raking procedure results in comparable CVs across the fifteen primary strata, it requires higher sample sizes in strata with smaller sales. The results of the c-scenario (Figure 3.3) are close to those obtained by the n-scenario (Figure 3.4), when $q = 0.1$.

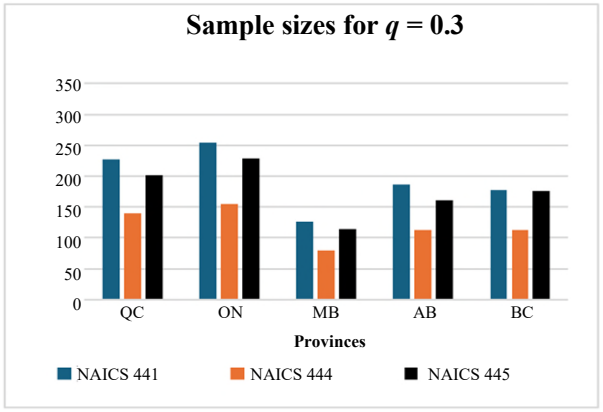
Figure 3.4 Primary strata sampling sizes (n_{gi}) obtained using various q values (0.0, 0.1, 0.3, 0.5, and 1.0) and corresponding coefficients of variation (c_{gi}), $g = 1,\ldots,5$ and $i = 1, 2, 3$

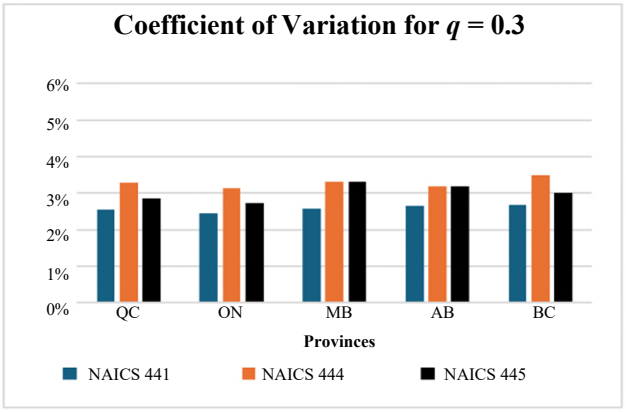












Note: NAICS = North American Industry Classification System.

Figure 3.4 (continued) Primary strata sampling sizes (n_{gi}) obtained using various q values (0.0, 0.1, 0.3, 0.5, and 1.0) and corresponding coefficients of variation (c_{gi}), $g = 1,\ldots,5$ and $i = 1, 2, 3$

Note: NAICS = North American Industry Classification System.

4. Sample rotation and coordination

4.1 Introduction

In this section, we review methods for initial sample selection, sample rotation, and the selection of new businesses (births) in business surveys. These methods are designed to ensure that the sample consistently reflects the structure of the population over time, allowing for accurate and reliable estimates. The key objectives of these methods are to reduce the response burden on businesses, especially smaller ones, and to improve the quality of estimates, such as changes and annual averages.

Sample rotation involves periodically replacing part of the sample while retaining a portion of it. This strategy helps to minimize the burden on businesses. A critical aspect of sample rotation is adhering to "time-in" and "time-out" requirements. These constraints are designed to ensure that businesses are not

overburdened by the survey while maintaining sample integrity. The time-in requirement refers to the minimum duration for which businesses should remain in the sample. For example, a typical time-in-period might be 12 months. The time-out requirement stipulates that businesses rotated out of the sample must not be re-selected for a specified period, typically 12 months. This helps prevent over-surveying, which can lead to response fatigue and potential data distortion.

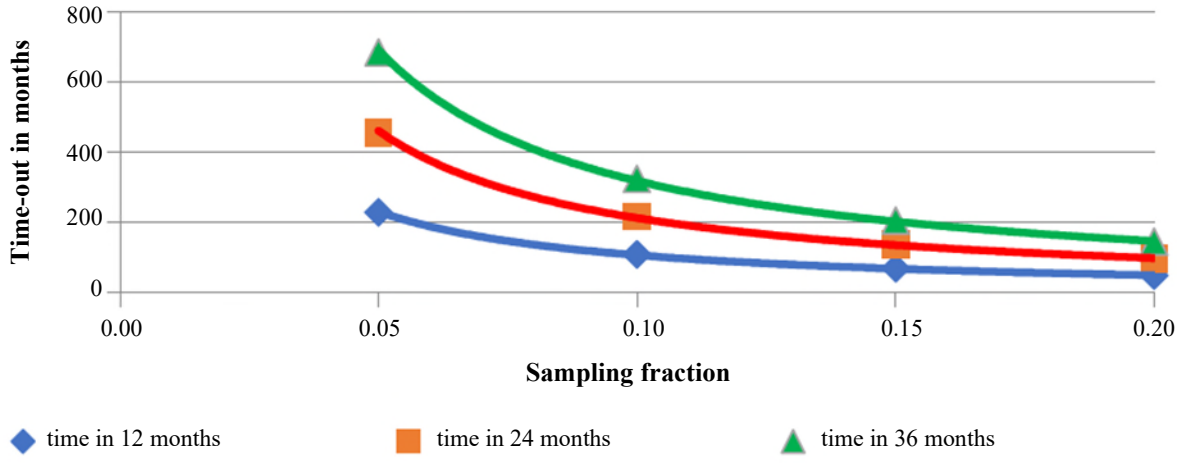
In practice, some units may need to remain in the sample for durations longer than the prescribed time-in-period to satisfy both constraints and preserve the integrity of the sampling design. While this may slightly increase the response burden for those businesses, it is generally more efficient than rotating them back into the sample within a short time.

For a given take-some size stratum U_h, let the sampling rate be f_h. Define $t_{h,\text{in}}$ and $t_{h,\text{out}}$ as the number of occasions that a unit will be in the sample and out of the sample, respectively. The following relationship holds $t_{h,\text{out}} = f_h^{-1}(1-f_h)\,t_{h,\text{in}}$. Figure 4.1 presents a graph where the x-axis represents sampling fractions, and the y-axis represents time-out. The graph includes several curves or data points corresponding to different values of time-in.

From Figure 4.1, we observe that:

1. There is a positive relationship between time-in and time-out. As the time allocated for the active process (time-in) increases, the time spent in the inactive phase (time-out) also increases.
2. For any fixed value of time-in, the time-out period tends to decrease exponentially as the sampling fraction increases. This suggests that higher sampling fractions lead to faster or more efficient transitions between active and inactive periods.
3. For a constant sampling fraction, as time-in increases, time-out also increases. This indicates that when more time is spent on the active process, there is a longer required time-out, regardless of the sampling fraction.

Figure 4.1 Relationship between time-in and time-out for given sampling rates



There are two primary types of rotation schemes used in business surveys: *survey-design-dependent* and *survey-design-independent* schemes.

1. A *survey-design-dependent* rotation scheme only allows for the rotation of sampled units within the strata of a single survey. However, it does not manage or control the overlap of sampled units across different surveys.
2. A *survey-design-independent* rotation scheme, on the other hand, supports both within-survey rotation and controlled sample overlap across multiple surveys. This controlled overlap is referred to as *sample coordination*, which is essential for reducing response burden and improving efficiency when multiple surveys are drawn from a common sampling frame.

There are several reasons for implementing sample coordination in business surveys. In some cases, the objective is to *reduce the response burden*, which is achieved through *negative coordination*–that is, designing the sample selection to avoid including the same business in multiple surveys within a short time frame. In other situations, the goal may be to *maximize sample overlap* between surveys, known as *positive coordination*. This approach enhances continuity during survey redesigns, reduces transition costs, and facilitates the use of data from one survey to support another. A *hybrid approach*, combining both positive and negative coordination, is also possible. For example, one might aim to maintain a high degree of sample overlap between successive occasions of a given monthly survey (positive coordination), while simultaneously limiting the time any unit remains in the sample to avoid overburdening respondents (negative coordination).

We next provide two examples of survey design-dependent rotation schemes to illustrate how these principles are applied in practice.

4.2 Survey design dependent rotation schemes

In the early to mid-1980s, Statistics Canada introduced rotation schemes for two of its flagship business surveys: the Survey of Employment, Payrolls and Hours (SEPH) and the Monthly Wholesale and Retail Trade Survey (MWRTS). The SEPH provides a monthly statistical portrait of employment earnings, the number of jobs (i.e., occupied positions), and hours worked, broken down by detailed industry classifications at the national, provincial, and territorial levels. The MWRTS, on the other hand, collects monthly data on retail and wholesale sales and inventories, disaggregated by both industry and geography. At the time of implementation, both surveys were based on the 1980 Standard Industrial Classification (SIC) system, which provided a consistent framework for categorizing business activity across Canada.

SEPH implemented a rotation method known as rotation group sampling, while MWRTS adopted a different approach called panel sampling. Despite employing different techniques, both surveys applied rotation strategies that aligned with the stratified design of their samples. Specifically, the rotation of sampled units was conducted within each secondary size stratum, maintaining the integrity of the stratification. These surveys continue to be operational. Recent documentation on SEPH is given in SEPH (2025).

MWRTS is currently run as two separate surveys, known as the Monthly Retail Trade Survey (see MRTS, 2025) and the Monthly Wholesale Trade Survey (see MWTS, 2025).

Below is a summary of the methodologies used in each rotation scheme for SEPH and MWRTS in the early nineties.

4.2.1 Rotation groups

SEPH developed its rotation scheme in the early 1980s. A brief overview of this scheme is as follows. The monthly sample consisted of 14 groups, numbered 0 through 13. Group 0 included the take-all units within a primary stratum. Groups 1 to 13 were labeled as rotation groups. The numbers 1 through 12 for the rotation groups corresponded to the month in which the units, other than births, rotated into the sample. For example, rotation group 1 primarily consisted of units that entered the sample in January, as well as any births, while rotation group 2 contained units that entered the sample in February, and so on. Rotation group 13 consisted of units that had been in the sample for 12 months: these were the oldest units in terms of time spent in the sample. These units were eligible to be rotated out. Each month, new births were randomly selected and allocated to the rotation groups. Details on the rotation method used for SEPH are provided in Şchiopu-Kratina and Srinath (1991).

4.2.2 Panel sampling

The Monthly Wholesale and Retail Trade Survey (MWRTS) used a rotation scheme based on panels. For this scheme, the population within each secondary size stratum was randomly divided into rotation groups, ensuring that the number of units in each group differed by no more than one unit. The total number of rotation groups was determined based on sampling fractions and time-in/time-out constraints. A simple random sample of rotation groups was selected, with the proportion of groups chosen aiming to match the desired sampling fraction for each stratum. The sample included all units from the selected rotation groups. Rotation occurred by adding new units (births) to the groups in a systematic manner while removing in-sample groups, and deaths were only removed if identified by an independent source or after a specified period. Sample updates, based on the methods of Kish and Scott (1971), aimed to maximize the overlap between the current and new samples, minimizing disruption, operational costs, and discontinuities in estimates. Details of the panel procedure are discussed in Hidiroglou, Choudhry and Lavallée (1991).

4.3 Survey design independent rotation schemes

A key element in coordinated sampling is the use of a Permanent Random Number (PRN). The PRN is a uniformly distributed random number in the range [0, 1], permanently assigned to each unit in the population. This number remains attached to the unit for as long as it exists and maintains a continuous and identifiable presence in the population frame.

PRN-based sampling was developed independently in the early 1970s by the Australian Bureau of Statistics and Statistics Sweden. It enables flexible and coordinated sample designs across time and surveys by using a consistent, reproducible method for sample selection.

Three commonly used sampling procedures that leverage the PRN are:

- *Poisson sampling*: Units are selected independently with inclusion probabilities defined as a function of the PRN.
- *Synchronized sampling*: Ensures consistent inclusion or exclusion of units across coordinated surveys by aligning selection intervals with the PRN.
- *Microstrata sampling*: Divides the PRN range into small strata (microstrata) to achieve controlled selection and rotation within strata.

4.3.1 Poisson sampling

An early use of the PRN was Poisson Sampling. Brewer, Early and Joyce (1972) suggested the following method of selecting a Poisson Sample as a way to handle both rotation and overlap.

For a population U of size N, assign a random number $r_k, k = 1,\ldots,N$ that is uniformly distributed in the interval [0, 1] to the kth unit. The kth unit is included in the initial sample if $r_k \in [0, \pi_k]$. Here π_k is the probability of inclusion in the sample given by $\pi_k = n x_k \Big/ \sum_{j=1}^{N} x_j$, where x_j is the measure of size associated with the jth unit in population U. The realized sample size n_s is random, with mean $E(n_s) = \sum_{k=1}^{N} \pi_k$ and variance $V(n) = \sum_{k=1}^{N} \pi_k (1-\pi_k)$. Rotation of the sample takes place by shifting the sampling window progressively by a rightward shift τ. At the second occasion, the kth unit is included in the sample if i. $r_k \in [\tau, \pi_k + \tau]$ or ii. $r_k \in [\tau, 1] \cup [0, \pi_k + \tau - 1]$, where $\pi_k + \tau > 1$. Its chief disadvantage is that the realized sample size is a random variable with a relatively high variance.

4.3.2 Synchronized sampling

The method of selection was adapted from the JALES method (Atmer, Thulin and Bäcklund, 1975). JALES stands for the initials of the authors who developed it, Johan Almer and Lars-Erik Sjoberg, both former employees at Statistics Sweden. A good description of the method is given in Brewer, Gross and Lee (2000) and McKenzie and Gross (2000). The Australian Bureau of Statistics introduced synchronized sampling in the early 1980s: see Hinde and Young (1984) for a detailed description of its computer application. This method can control rotation at the survey level, allows for different stratification in surveys, and can control the overlap (coordination) of sampled units between surveys. Coordination is either negative or positive. For negative coordination, the expected overlap is minimized. For positive coordination, the expected sample overlap is maximized.

Rotation takes place as follows. For the first selection, a provisional start point, s_0, is set, either purposively or randomly, between 0 and 1. Given the required sample size n, the first n units located at or to the right of s_0 are selected. They are described by the interval $[s_1, e_1)$ (closed on the left and open on the

right), where the start point, s_1, is set as the first unit at or to the right of s_0 and the endpoint, e_1, is set as the $(n+1)$th unit at or to the right of s_0. For the next selection, only the start and end points s_1 and e_1 need to be remembered, and the number of units from the new population in the interval $[s_1, e_1)$ is compared with the desired sample size n. If the correct number of units is in the interval (i.e., there have been zero net births or deaths in the selection interval), then the new start and end points, s_2 and e_2, generally remain at s_1 and e_1. The exception is when the unit on s_1 or e_1 dies, in which case the next unit to the right becomes the boundary point. The overlap between surveys is minimized as follows. When there is rotation, the position of the selection interval changes from one cycle of a survey to another. To control overlap between surveys, each stratum of each survey is given a fixed interval, the survey range, and the selection interval is forced to remain within the survey range.

The survey range must be large enough to hold the selection interval, plus enough unselected units to permit worthwhile rotation to occur. The overlap between surveys is maximized as follows. If two surveys have the same stratification, the overlap between them is maximized by using one selection interval for the smaller survey and another disjoint interval for the extra sample for the larger survey. When the surveys have different stratifications, positive coordination is only achieved by manual choice of the selection interval.

4.3.3 Microstrata

The Microstrata method, as detailed in Rivière (2001), is a robust approach to sample coordination designed to manage response burden and optimize overlap control across business surveys. This method relies on systematic permutations of initially assigned random numbers within carefully constructed subgroups of the population, referred to as *microstrata*. These permutations are strategically used for multiple coordination objectives, including the smoothing of response burden over time, the minimization of sample overlap between surveys, and the maintenance of updated panels.

The concept of microstratification was developed in 1998 under the auspices of Eurostat's SUPCOM project (Rivière, 1998) on sample coordination, with its first software implementation realized through the SALOMON system in 1999. Further enhancements to the methodology culminated in the release of MICROSTRAT in 2001, which offered expanded functionality and greater flexibility for practical implementation. A good discussion of this procedure is given in Nedyalkova, Pea and Tillé (2008).

Under this method, each unit in the sampling frame (e.g., the BR) that is within scope is assigned a fixed random number drawn from a uniform distribution. This number serves as the basis for sample selection. Additionally, each unit receives a response-burden coefficient, which accumulates over time based on its inclusion in one or more surveys. Although the random numbers themselves remain static, permutations are allowed across units, conditioned on their respective burden coefficients. These permutations occur within *microstrata*, which are defined as the finest possible cross-classification of units that allows sorting by increasing response burden without introducing selection bias.

The key statistical properties of the Microstrata method include:

- *Preservation of randomness*: The fixed random numbers retain their uniform distribution post-permutation, ensuring selection probabilities remain valid.
- *Bias avoidance*: The permutations within microstrata are constructed to maintain design-unbiasedness in survey estimates.
- *Automatic handling of dynamics*: The method gracefully accommodates unit births, deaths, and strata changes, offering a high degree of adaptability to frame updates and population shifts.
- *Stratification flexibility*: There are no strict constraints on the number of strata or the rotation rates, which enhances its general applicability across various business surveys.

Despite its advantages, a notable limitation is the potential formation of microstrata that are too small, which can impair the effectiveness of coordination. However, this can be mitigated through alternative sorting schemes or aggregation rules that preserve coordination goals without compromising the statistical integrity of the strata.

In Canada, the Survey of Employment, Payrolls and Hours (SEPH) adopted the Microstrata method for its sample rotation procedures starting in 2010, as documented in Landry (2011).

4.3.4 Comparing survey design independent rotation to survey design dependent rotation schemes

Srinath and Carpenter (1995) conducted a comparative study of several rotation schemes over fifty consecutive periods, evaluating both panel sampling (a survey design–dependent rotation scheme) and Bernoulli sampling (a survey design–independent rotation scheme). Their simulation model accounted for dynamic population changes by incorporating both *births* (new units entering the population) and deaths (units exiting the population) at each time point, providing a realistic assessment of the rotation methods' performance over time.

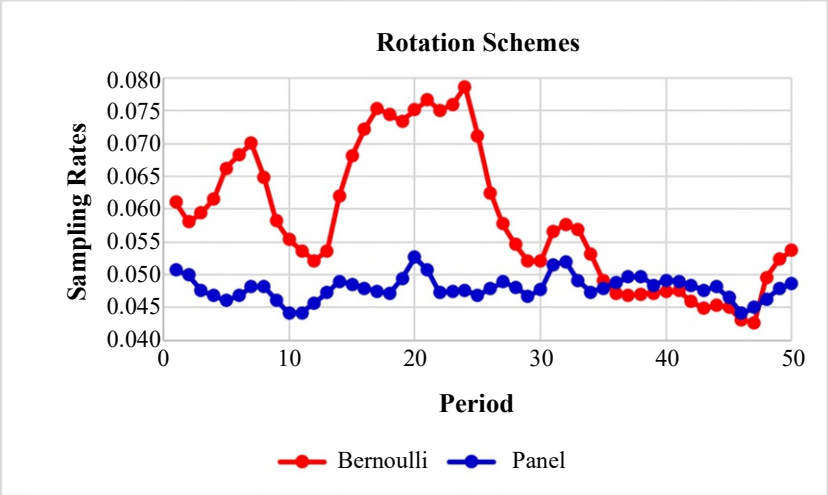
The initial sample was drawn from a population of size 500. On each subsequent occasion, the number of births and deaths was assumed to follow Poisson distributions with means of 20 and 15, respectively. Additionally, they assumed a 70% probability of identifying a death through an external source.

The results are illustrated in Figure 4.2, which corresponds to Figure 10.3(A) in Srinath and Carpenter (1995). The figure demonstrates that Bernoulli sampling exhibits the greatest variation in sampling fractions over time, highlighting one of its key limitations in maintaining sample stability.

The large differences in expected sampling rates between panel sampling and Bernoulli sampling have important implications for the estimation of totals. Panel sampling tends to produce a more continuous time series than Bernoulli sampling, provided that births and deaths do not significantly disrupt the balance of units between panels. However, if an imbalance arises in the number of units across panels, even composite estimation may fail to yield smooth estimates across survey occasions.

The Monthly Wholesale and Retail Trade Survey, redesigned in the mid-1980s, initially adopted a panel sampling design. However, the rotation of units was halted early in its implementation because the month-to-month estimates exhibited excessive variability. This instability was primarily due to poor stratification of the smaller units within the sampling frame.

Figure 4.2 Comparing sampling rates for two rotation schemes



In contrast, the U.S. Census Bureau's monthly retail and wholesale trade surveys employed a design where sampled firms were divided into three panels, each reporting every third month. This approach provided overlapping data for the current and previous months, and composite estimation was used to reduce the variance of both levels and changes. Despite these efforts, rotation for these surveys was discontinued in 1997 in favor of a design without panels (see Cantwell and Caldwell, 1998 for details).

These surveys publish estimates of totals. The Survey of Employment, Payrolls, and Hours (SEPH), which has used rotation since the early 1980s and continues to do so with the MICROSTRAT method, shows changes in estimates for successive occasions – such as average weekly earnings and payroll employment – that remain within acceptable bounds. One reason for this is that these variables exhibit much less skewness compared to the retail and wholesale trade variables, leading to more stable and reliable estimates.

5. Elimination of dead units

Deaths accumulate over time in a periodic survey, particularly when it is conducted frequently (e.g., monthly or quarterly) and the sample is not independently selected at each cycle. While leaving the dead units in the sample does not bias the estimates, it does inflate the variance over time, since these units are coded as zeroes. To mitigate this, it is important to remove dead units both from within the sample and from the population frame. However, the removal must be done carefully to avoid introducing bias into the

estimates over time. The initial sampling weights, based on the original sampling fractions, must be adjusted to account for the impact of removing dead units.

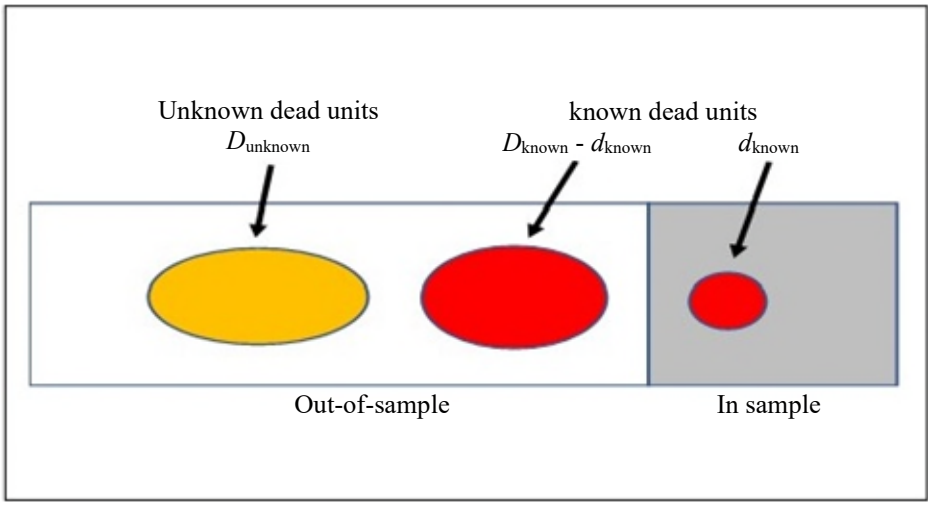
In what follows, the procedure is described at the level of the secondary stratum U, which is defined based on unit size. Suppose that on a given survey occasion, stratum U has N population units, from which n units are sampled using simple random sampling without replacement. The resulting sampling rate is $f = n/N$, and an initial sampling weight $w = 1/f$ is associated with each sampled unit. This weight remains fixed throughout the life of the survey.

Suppose that D_{known} and d_{known} represent the *known* number of dead units in stratum U and sample s, respectively. The difference, $D_{\text{known}} - d_{\text{known}}$, corresponds to the *known* dead units that are out-of-sample $U \setminus s$. These *known* dead units have been identified either through the survey sample or via an administrative source that updates the population register. In contrast, *unknown* dead units refer to out-of-sample units that are deceased but have not yet been identified by the administrative source. Let D_{unknown} denote the number of *unknown* out-of-sample deaths.

Figure 5.1 illustrates the problem: how does one determine the number of in-sample and out-of-sample dead units to retain so that they accurately represent the unknown number of out-of-sample dead units?

The number of estimated known and unknown dead units in the population is wd_{known}. This implies that the number of estimated known and unknown dead units out of the sample is $(w-1)\,d_{\text{known}}$. It follows that the estimated number of unknown dead units out of the sample is $\hat{D}_{\text{unknown}} = (w-1)\,d_{\text{known}} - (D_{\text{known}} - d_{\text{known}})$. This difference can be negative, zero, or positive. If $\hat{D}_{\text{unknown}} \leq 0$, no in-sample or out-of-sample deaths are removed. If $\hat{D}_{\text{unknown}} > 0$ we can remove some in-sample *known* dead units, as well as some out-of-sample *known* dead units, so that the number of remaining in-sample dead units represents the out-of-sample *unknown* deaths.

Figure 5.1 Known and unknown deaths



Let x be the number of dead units to be left in the sample. These units represent $(w-1)\,x = \hat{D}_{\text{unknown}}$ *unknown deaths* in the population.

Given that $\hat{D}_{\text{unknown}} = (w-1)\,d_{\text{known}} - (D_{\text{known}} - d_{\text{known}})$, we have

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

Solving equation (5.1) for x, the number of in-sample deaths that must be kept in the sample s to represent unknown out-of-sample deaths is

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

where $d^{*}_{\text{known}} = (D_{\text{known}} - d_{\text{known}}) \,/\, (w-1)$. Here d^{*}_{known} is the number of dead units to be removed from the sample.

Correspondingly, $w\,d^{*}_{\text{known}}$ estimates the total number of deaths removed from the d_{known} dead units in the sample. These deaths cannot be removed from the unknown dead units D_{unknown}, so they must be removed from the known dead units. The estimated number of known dead units is $\hat{D}^{*}_{\text{known}} = w\,d^{*}_{\text{known}}$. It follows that the number of dead units that should be removed from the out-of-sample units $(D_{\text{known}} - d_{\text{known}})$ should be $\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}}$. Note that the number of units to be removed from the sample (d^{*}_{known}) and out-of-sample $(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})$ are most likely real numbers; therefore, they need to be converted into integers. To this end, we denote the integer portion of d^{*}_{known} as $d^{*}_{\text{known,int}}$ where $d^{*}_{\text{known,int}} \le d^{*}_{\text{known}}$. Hence, the number of units to be removed from the in-sample dead units will be

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

where int (a) is the integer portion of the real number a.

The number of dead units to be retained in the sample is $d_{\text{known}} - d^{*}_{\text{known,int}}$. Correspondingly, the number of out-of-sample known dead units that are to be removed from the $D_{\text{known}} - d_{\text{known}}$ out-of-sample dead units is:

$$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}} = \text{int}\left\{(w-1)\,d^{*}_{\text{known}} + 0.5\right\}. \tag{5.3}$$

This results in the three scenarios for removing both in-sample and out-of-sample dead units are summarized in Table 5.1.

Table 5.1
Outcomes for removing deaths

Case	Condition	In-sample dead units removed	Out-of-sample dead units removed
1.	$d^{*}_{\text{known,int}} = 0$	No deaths	No deaths removed
2.	$d^{*}_{\text{known,int}} = d_{\text{known}}$	All deaths	$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}}$ deaths removed
3.	$0 < d^{*}_{\text{known,int}} < d_{\text{known}}$	$d^{*}_{\text{known,int}}$ deaths	$(\hat{D}^{*}_{\text{known}} - d^{*}_{\text{known}})_{\text{int}}$ deaths removed

When the dead units are removed from the sample and out-of-sample, we are left with $(N - d^*_{\text{known,int}}) - (\hat{D}^*_{\text{known}} - d^*_{\text{known}})_{\text{int}}$ units in the stratum population and $n - d^*_{\text{known,int}}$ units in the sample. The removal of the dead units should result in minor differences between the previous weight w and the new weight.

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

Denote the universe and sample after death removal as U_c and s_c respectively. If the estimator is a total that uses a variable of interest y, we could estimate the total for the population by $\hat{Y}^*_{\text{HT}} = \sum_{i \in s_c} w y_i$.

A better estimator that reflects that the death removal has affected both the population and sample counts is the Hájek estimator, given by

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

6. Summary

Business surveys play a crucial role in collecting data that support informed decision-making in both the public and private sectors. The quality of these surveys depends on several key factors that include: i. a comprehensive, unduplicated, and up-to-date BR; ii. A sampling design that reflects the real-world structure of businesses and enables optimal sample selection; iii. The integration with administrative data sources to enhance efficiency and data quality; iv. A generic processing system that supports all stages of business survey operations, including data collection, editing, estimation, sampling, and imputation; v. Effective feedback mechanisms between the BR and the surveys drawn from it; vi. Strategies to reduce response burden, such as well-designed questionnaires, sample rotation within a given survey, and coordination across surveys to limit the repeated selection of the same businesses; and vii. The removal of dead units ensures accurate variance estimation.

This paper has mainly focused on: i. the construction, structure, and ongoing maintenance of the business register; ii. Methods for determining and allocating sample size; iii. Techniques for sample rotation and coordination; and iv. A procedure for the elimination of dead units.

These methodological improvements have evolved and have enabled Statistics Canada to establish a robust and reliable business statistics program.

Acknowledgements

The author thanks the editor and the referees for their excellent comments that improved the paper.

Appendix A

Multivariate extension to Bankier (1988)

Define a multivariate vector $\mathbf{y}_k = (y_{k1}, y_{k2}, \ldots, y_{kJ})$ for $k = 1, \ldots, N,$ with dimension J. The problem is to allocate a sample size n selected from a population U of size N to H strata $(U_h, h = 1, \ldots, H)$. Extending Bankier (1988), we minimize the function F_2

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

under the constraint $\sum_{h=1}^{H} n_h \kappa_h = \mathrm{K}, h = 1, \ldots, H,$ where

- κ_h is the cost of enumeration per unit in stratum h;
- n_h is the sample size allocated to stratum h;
- K is the total allotted cost;
- X_h^q is the importance of stratum $h,$ and q is a constant in the range $0 \le q \le 1$;
- r_j is the relative importance of a variable y_j regardless of the stratum;
- $Y_{hj} = \sum_{k \in U_h} y_{kj}$ is the population total of the variable y_j in stratum h;
- $\hat{Y}_{hj} = N_h \sum_{k \in s_h} y_{kj} / n_h$ estimates the total Y_{hj}.

The variance $V(\hat{Y}_{hj})$ is

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

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

F_2 is minimized using the Lagrangian function $F_3,$ where

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

with $A_{hj}^2 = N_h^2 S_{hj}^2 / Y_{hj}^2$ and λ is the Lagrange multiplier. F_3 is differentiated with respect to n_h $(h = 1, \ldots, H)$ and λ. Setting the resulting equations to zero, we obtain

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

and

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

Solving for n_h from (A.1) yields:

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

Substituting n_h from (A.3) into (A.2), we obtain

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

From equation (A.4), we have that

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

Substituting $1/\sqrt{\lambda}$ from (A.5) into (A.3), we obtain that

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

If $J = 1$, we obtain the unidimensional version as

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

If $\kappa_h = \kappa$ for $h = 1, \ldots, H$, $K = n$, n_h is

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

This corresponds to equation (3.2): the gi terms in that equation are replaced by h.

Appendix B

Equivalence between the c and n scenarios

Let the continuous version of the discrete distribution of the x-variable be $g(x)$. The continuous version of equation (3.8) to minimize the overall sample size n for a specified coefficient of variation c is given by:

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

where

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

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

The terms $\omega_h\ \mu_h,$ and μ in $A, B, D,$ and $F,$ are the continuous analogs of $W_h\ \bar{X}_h\ \sigma_h^2,$ and $\bar{X}$ respectively. Differentiating (B.1) with respect to $b_h, h = 1,2,\ldots,H-1,$ and equating the resulting equations to 0, we obtain:

$$F(A_h^{(1)}B + AB_h^{(1)}) - ABD_h^{(1)} = 0 \text{ for } h = 1,...,H-2 \tag{B.2}$$

and

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

where $A_h^{(1)}$, $B_h^{(1)}$, and $D_h^{(1)}$ are the derivatives of $A, B,$ and D with respect to b_h; $h = H-1,$ and $\omega_H^{(1)}$ is the derivative of ω_L with respect to b_h for $h = H-1.$

The continuous version of equation (3.7) to minimize $V(\hat{\bar{X}})$ for a specified sample size n is given by:

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

Differentiating (B.4) with respect to $b_h, h = 1,2,\ldots,H-1,$ and equating the resulting equation to 0, we obtain:

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

and

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

We now show that equations (B.2) and (B.5) are equivalent for $h = 1,\ldots,H-2.$ From equation (B.1), we have that

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

Substituting (A.7) into (A.5), we have

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

Hence, the result follows. Using a similar argument, it can be shown that equations (A.3) and (A.6) are equivalent for $h = H - 1$.

References

Atmer, J.G., Thulin, G. and 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. and Rivest, L.-P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77, 331-344.

Baillargeon, S., Rivest, L.-P. and Ferland, M. (2007). Stratification en enquêtes entreprises : Une revue et quelques avancées. *Proceedings of the Survey Methods Section*, Statistical Society of 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. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 81-86. https://ssc.ca/sites/default/files/survey/documents/SSC2001_H_Berard.pdf.

Brewer, K.R.W., Early, L.J. and 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. and Lee, G.F. (2000). PRN Sampling: The Australian Experience. Proceedings: *Invited Papers, IASS Topics*, Helsinki August 10-18, 1999, 155-163. https://ww2.amstat.org/meetings/ices/2000/proceedings/S31.pdf.

Brodeur, M., Koumanakos, P., Leduc, J., Rancourt, É. and Wilson, K. (2005). The integrated approach to economic surveys in Canada. Statistics Canada, Catalogue No. 68-514-XWE. https://www150.statcan.gc.ca/n1/pub/68-514-x/68-514-x2006001-eng.htm.

Cantwell, P. and 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. In *Business Survey Methods*, (Eds., B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. College and 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. and 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*, July 2023. https://www.isi-next.org/media/abstracts/ottawa-2023_8dcef16b7fa9e697566aa54408a57532.pdf.

Deming, W.E. and 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.

Ferland, M. and Batten, D. (2007). Sampling for the MWRTS Redesign. Working Paper No. BSMD-2007-002E, Methodology Branch, Ottawa: Statistics Canada.

Hidiroglou, M.A. (1986a). Wholesale/Retail Trade Surveys Redesign. Working Paper No. BSMD-87-017E, Methodology Branch, Ottawa: Statistics 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. and Lavallée, P. (1991). A sampling and estimation methodology for sub-annual business surveys. *Survey Methodology*, 17(2), 195-210. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14502-eng.pdf.

Hidiroglou, M.A. and Kozak, M. (2018). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*, 86(1), 87-105.

Hidiroglou, M.A. and Laniel, N. (2001). Sampling and estimation issues for annual and sub-annual Canadian business surveys. *International Statistical Review*, 69(3), 487-504.

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

Hinde, R. and Young, D. (1984). Synchronized Sampling and Overlap Control Manual. Unpublished report, Australian Bureau of Statistics.

Kirkendall, N.K., White Jr, G.D., Citro, C.F. and 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. and 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. and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14(1), 33-43. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1988001/article/14602-eng.pdf.

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

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

MRTS (2025). Monthly Retail Trade Survey, Detailed information for February 2025. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1570101.

MWTS (2025). Monthly Wholesale Trade Survey, Detailed information for February 2025. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1566980.

Nedyalkova, D., Pea, J. and Tillé, Y. (2008). Sampling procedures for coordinating stratified samples: Methods based on microstrata. *International Statistical Review*, 76(3), 368-386.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28(2), 191-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2002002/article/6432-eng.pdf.

Rivest, L.-P. and 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). Coordinating samples using the microstrata methodology. Proceedings: *Symposium 2001, Achieving data quality in a statistical agency: A methodological perspective*. Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2001001/session8/6252-eng.pdf.

Şchiopu-Kratina, I. and Srinath, K.P. (1991). Sample rotation and estimation in the survey of employment, payrolls and hours. *Survey Methodology*, 17(1), 79-90. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991001/article/14520-eng.pdf.

SEPH (2025). Survey of Employment, Payrolls, and Hours, Detailed information for February 2025. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=2612.

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

Statistics Canada (1997). Project to Improve Provincial Economic Statistics. https://publications.gc.ca/collections/collection_2017/statcan/11-204/CS11-204-1997-eng.pdf.

St-Louis, G. (2008). The evolution of administrative data use for the Canadian Business Register (BR). June 2008 Shanghai Conference. 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. and 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.