Sampling for Business Surveys at Statistics Canada

M.A. Hidiroglou
October 29, 2025

Outline

1. **Business Register**
2. **Sample size determination and allocation**
3. **Sampling from the Business Register**
4. **Removal of dead units**
5. **Concluding Remarks**

1. Business Register

- **Some History**
- **1970s:** Initial creation
  - Early challenges: incomplete coverage and fragmented data
- **Redesigns**
  - **1984-1988:** Business Survey Redesign Project → integrated administrative data, modular system. First survey to use it Monthly Retail and Wholesale trade Surveys
  - **1997:** Integration of Business Number (BN) → improved linkage and data quality
  - **Late 1990s:** Unified Enterprise Survey (UES) → unified annual surveys under one framework
  - **2010:** Integrated Business Statistics Program (IBSP) → standardized and efficient survey operations

1. Business Register

Membership

- Sampling frame for economic surveys and includes corporations, non-profits, and government entities.
- A business is part of the BR if it remits payroll deductions, earns over $30,000 annually, or files a corporate tax return.
- BR built from payroll, GST, and income tax data, all linked by the **Business Number (BN)** to ensure consistency and avoid duplication.

1. Business Register

Classification and Updates

Each unit is classified by:

- Industry → NAICS (North American Industry Classification System)
- Geography → Provincial and regional codes

Continuous updates:

- Business births and deaths
- Structural changes in complex enterprises
- Revisions from administrative data and surveys

1. Business Register

Sampling

- Sampling from the **Business Register (BR)** involves defining the target universe using its hierarchy of statistical units (enterprise, company, establishment, location) and then selecting a sampling unit at the same or a more detailed level.
- The choice of sampling unit impacts data collection, estimation, survey logistics, and linkage with administrative data.

1. Business Register

Statistical Unit	Definition	Example
Enterprise	Legal entity with consolidated financial data	Quarterly Survey of Financial Statements
Company	Measures operating profit and capital employed	Energy Research and Development Expenditures
Establishment	Single production unit with homogeneous activity	Business Payrolls Survey
Location	Specific physical site	Job Vacancy and Wage Survey

2. Sample size determination and allocation

Stratify Business population by

- Geography, industry
- Size within geography and industry

Parameter of interest: For each primary stratum gi estimate totals means $\bar{Y}_{gi}$ or totals Y_{gi}

Require auxiliary variable X

2. Sample size determination and allocation

Stratification implemented in two stages by specifying primary and secondary strata

Primary strata *gi*: geography (*g*) by industry (*i*)

- Some complex businesses included with certainty

Secondary strata *h*: size groups within *gi*

- Take-all stratum (TA): Largest units sampled with certainty
- Take-some strata (TS): Smaller units sampled using simple random sampling
- Take-none stratum (TN): Smallest units are not sampled

2. Sample size determination and allocation

Primary strata: Two scenarios for allocating units

- ***c*-scenario: Choose the overall level of precision level *c***
- ***n*-scenario: Choose total sample size *n* for given budget**

Secondary strata: For each scenario, and each primary stratum

- Choose allocation procedure: Neyman, *X*-proportional
- Compute size boundaries and sample sizes

2. Sample size determination and allocation

Primary strata: ***c-scenario***

- Specify global coefficient of variation c
- Denote as gi primary stratum, g=1,…,G and i=1,…, I
- Associated total : X_{gi}
- Marginal totals: $X_{g.} = \sum_{i=1}^{I} X_{gi}$, $X_{.i} = \sum_{g=1}^{G} X_{gi}$,
- Overall total: $X_{..} = \sum_{g=1}^{G} \sum_{i=1}^{I} X_{gi}$

2. Sample size determination and allocation

Primary strata: *c-scenario*

- Compute marginal CVS
  - Geography: $c_{g.} = c\frac{X_{..}}{\sqrt{\sum_{g=1}^{G} X_{g.}^2}}, g = 1, \ldots, G$
  - Industry: $c_{.i} = \mathrm{c}\frac{X_{..}}{\sqrt{\sum_{i=1}^{I} X_{.i}^2}}, i = 1, \ldots, I$
- Set initial cv for *gi-th* primary stratum to

$$c_{gi}^{(0)} = 0.5(c_{g.} + c_{.i})$$

2. Sample size determination and allocation

Primary strata: *c-scenario*

- Iterate: $r=1,\ldots,R$

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

- Convergence after $R=5$ iterations
- Final coefficients of variation : $c_{gi}^{(f)}$

2. Sample size determination and allocation

Secondary strata: *c-scenario*

- Stratify each primary stratum into secondary strata
- Minimize primary sample size n_{gi} given
  - Coefficient of variation for *gi-th* primary stratum: $c_{gi}^{(f)}$
  - Allocation scheme of n_{gi} units to secondary strata:

    X-proportional, Neyman allocation
    - Obtain boundaries and associated secondary stratum sizes

***R* program** *:* Rivest and Baillargeon (2022)

2. Sample size determination and allocation

Primary strata: *n-scenario*

Bankier's (1988) method. Overall sample size *n* optimally allocated to primary strata

- Given $\sum_{g=1}^{G}\sum_{i=1}^{I} n_{gi} = n$, minimize

$$F = \sum_{g=1}^{G}\sum_{i=1}^{I}\left(X_{gi}^{q}\mathrm{CV}\left(\hat{Y}_{gi}\right)\right)^{2}, 0 \le q \le 1$$

- y: variable of interest
- *x:* auxiliary variable

2. Sample size determination and allocation

Primary strata: *n-scenario*

Sample size for gi^{th} primary stratum

$$n_{gi} = n \frac{\frac{S_{gi}(y) X_{gi}^{q}}{\bar{Y}_{gi}}}{\sum_{g=1}^{G} \sum_{i=1}^{I} \frac{S_{gi}(y) X_{gi}^{q}}{\bar{Y}_{gi}}}, 0 \leq q \leq 1$$

$S_{gi}(y)$: standard deviation of the y's

$\bar{Y}_{gi}$: population mean of the *y*'s

X_{gi}: population total of the *x*'s

2. Sample size determination and allocation

Primary strata: *n-scenario*

$S_{gi}(y)/\bar{Y}_{gi}$: not known

Assume $\frac{S_{gi}(y)}{\bar{Y}_{gi}} \cong a$ for all primary strata and fpc ignored

Then

$$n_{gi} \cong n \frac{X_{gi}^{q}}{\sum_{g=1}^{G} \sum_{i=1}^{I} X_{gi}^{q}}$$

2. Sample size determination and allocation

Primary strata: *n-scenario*

***q*=0 corresponds to** equal allocation

$$n_{gi} = n / GI$$

Properties

1. CVs for primary strata almost equal
2. CVs across primary strata can be large

2. Sample size determination and allocation

Primary strata: *n-scenario*

***q*=1 corresponds to** Neyman allocation

$$n_{gi} = n \frac{N_{gi} S_{gi}(x)}{\sum_{g=1}^{G} \sum_{i=1}^{I} N_{gi} S_{gi}(x)}$$

- $S_{gi}(x)$ the standard standard deviation of the *x*'s.
- **Properties**
  1. CV across primary strata is minimized.
  2. CVs at the primary stratum level can be large

Bankier (1988) recommended q =0.3 or 0.5 as a compromise

2. Sample size determination and allocation

Secondary strata: ***n-scenario***

- Stratify each primary stratum into secondary strata
  - Take-all, take-some, and take none
- Boundaries and sample sizes
  - Choose allocation procedure to secondary strata: Neyman, *X* proportional, power allocation
  - Minimize variance of estimated mean for each ***gi-th*** primary stratum: $V(\hat{\bar{X}}_{gi})$

2. Sample size determination and allocation

Setup: Comparison between *c* and *n* scenarios

- Used a simulated data set of 2,000 establishments of the Monthly Retail Trade Survey: Rivest and Baillargeon (2022)
- Fitted data to a gamma distribution
- Generated sale population values for establishments using predicted gamma values
  - Counts and sales made to agree with the June 2024 Monthly Retail Trade Survey tables published by Statistics Canada
  - Primary strata: five provinces and three industry groups

2. Sample size determination and allocation

Primary Strata

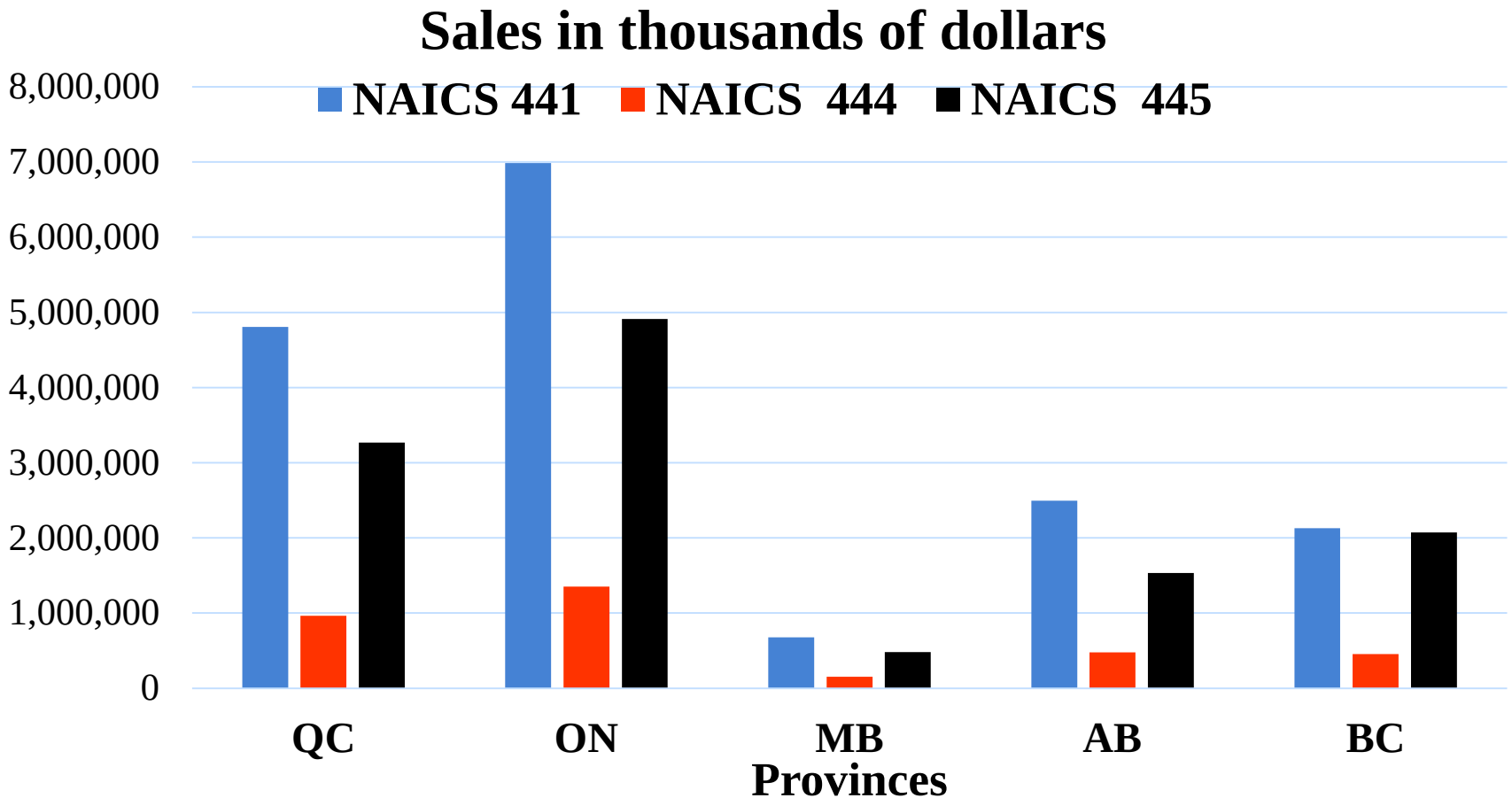
- **Provinces**: Quebec, Ontario, Manitoba, Alberta, British Colombia
- **Industries**

  NAICS 441: Motor Vehicle and Parts Dealers

  NAICS 444: Building Material and Garden Equipment and Supplies Dealers

  NAICS 445: Food and Beverage Stores

2. Sample size determination and allocation



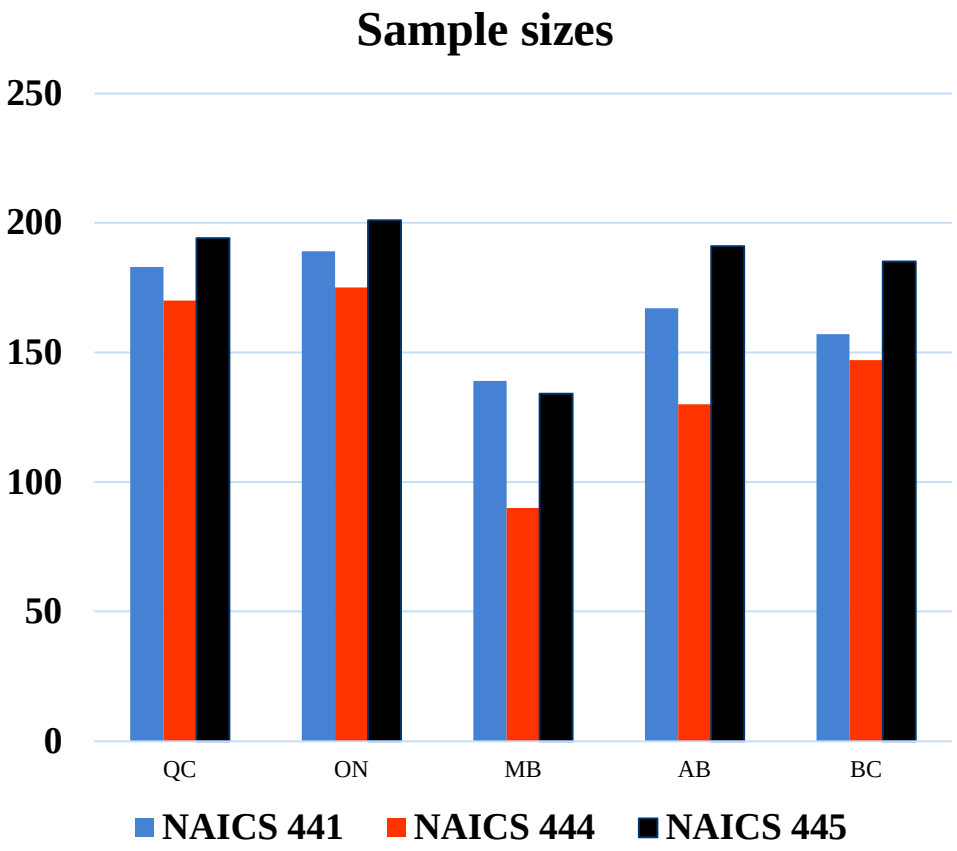
2. Sample size determination and allocation

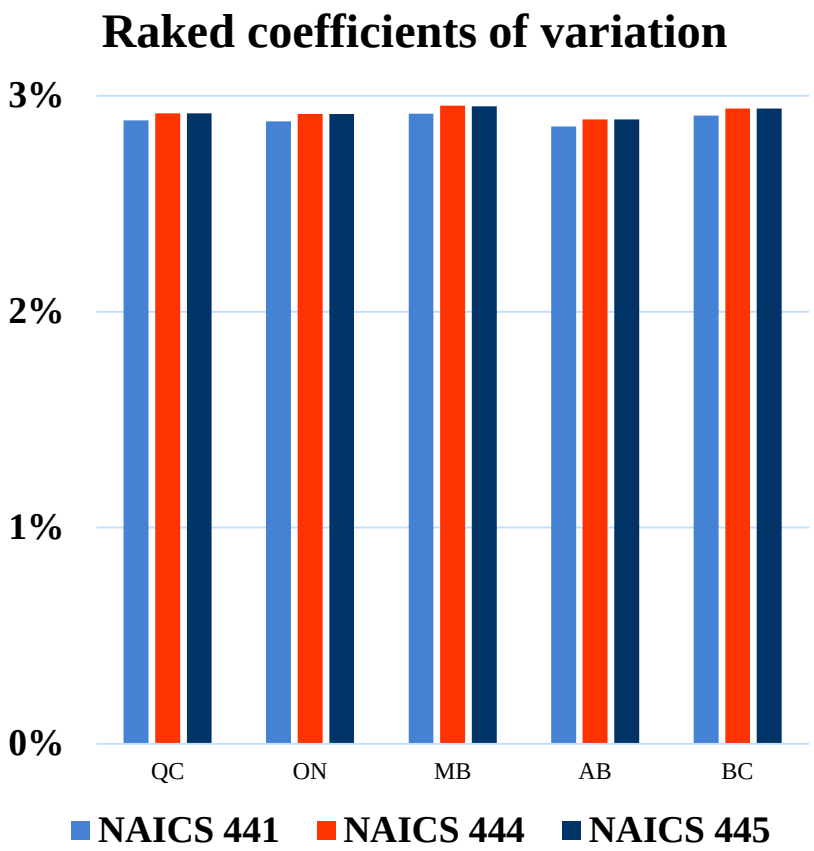
Steps for *c*-scenario

1. Overall coefficient of variation: 1%
2. Compute raked coefficients of variation for each primary stratum
3. Compute sample sizes for each secondary stratum given specified take-all stratum with two take-some strata using Neyman allocation.

***R* program** *:* Rivest and Baillargeon (2022)

Results for *c*-scenario





Summary *c*-scenario

1. Raked CVs are stable around 2.9%, indicating consistent precision.
2. Larger sample sizes correspond to slightly lower CVs.
3. Ontario and Quebec show the largest samples and lowest CVs.
4. Manitoba has smaller samples and higher CVs.
5. NAICS 444 shows the most variation across provinces
6. $CV \propto 1/\sqrt{n}$

2. Sample size determination and allocation

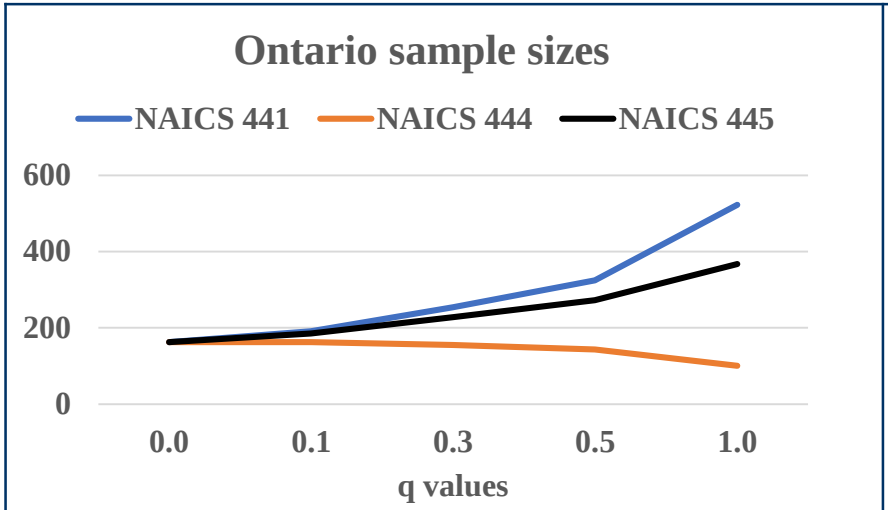
Steps for *n*-scenario

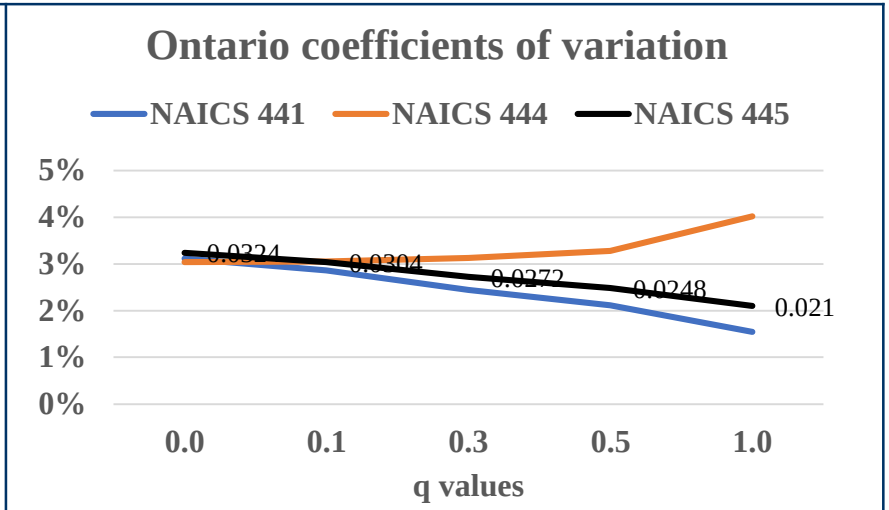
- Used overall sample size from *c*-scenario: n=2,452
- Allocated n to the primary strata using Bankier's (1988) power allocation

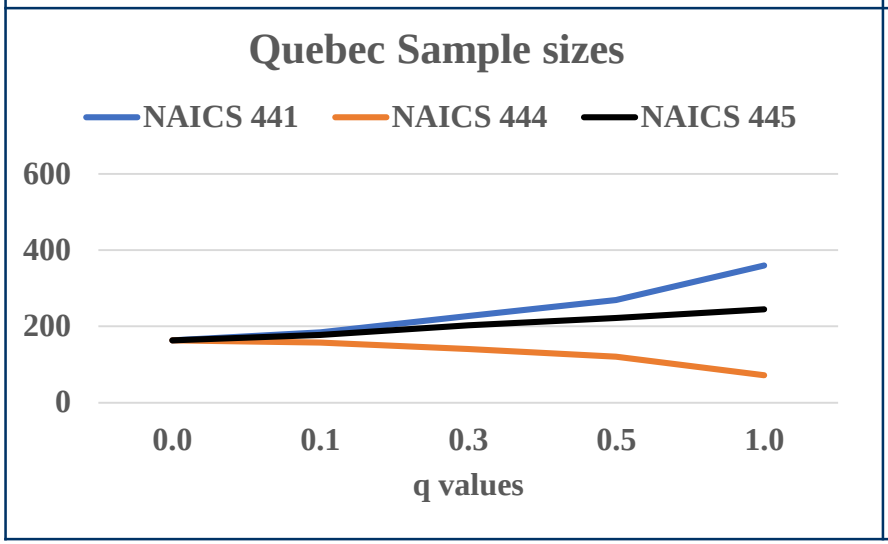
$$n_{gi} = n \frac{X_{gi}^{q}}{\sum_{g=1}^{G} \sum_{i=1}^{I} X_{gi}^{q}}$$

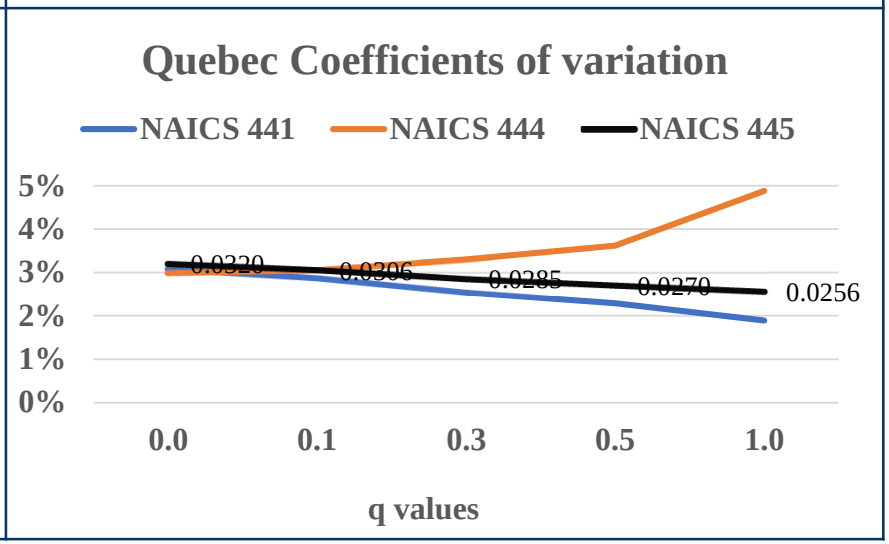
- q=(0.0,0.1, 0.3, 0.5, and 1.0)
- Computed sample sizes for each secondary stratum gi, given a take- all stratum and two take-some strata

Results for *n*-scenario

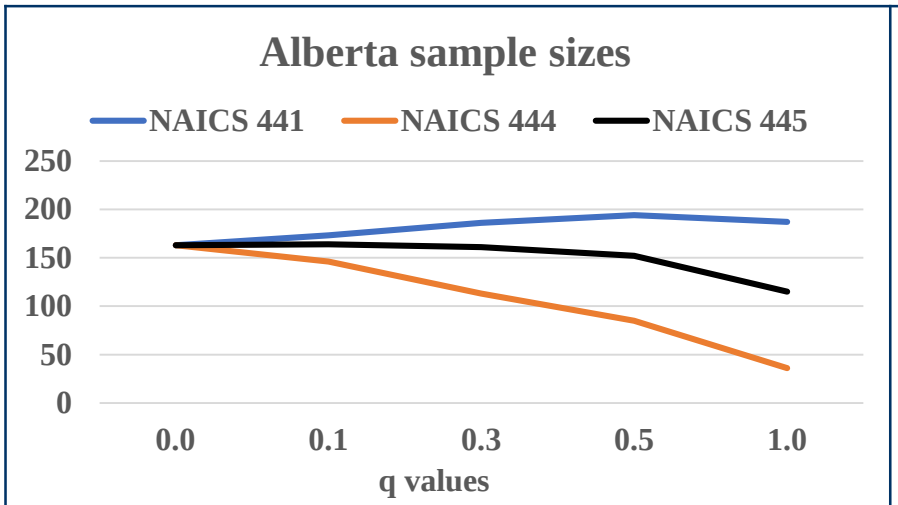


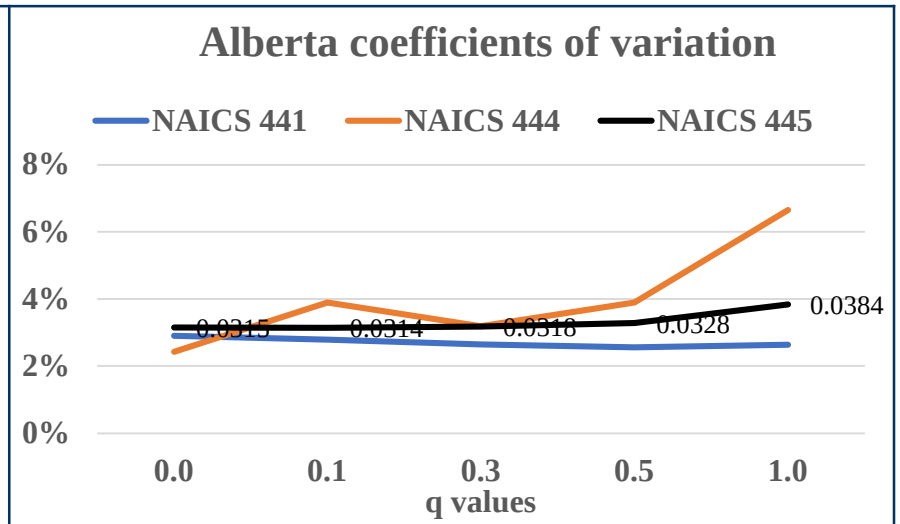


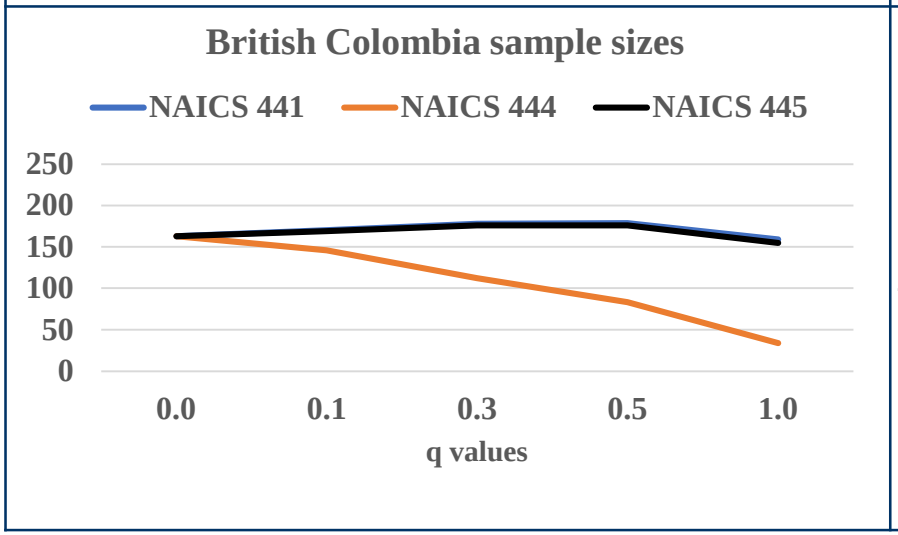


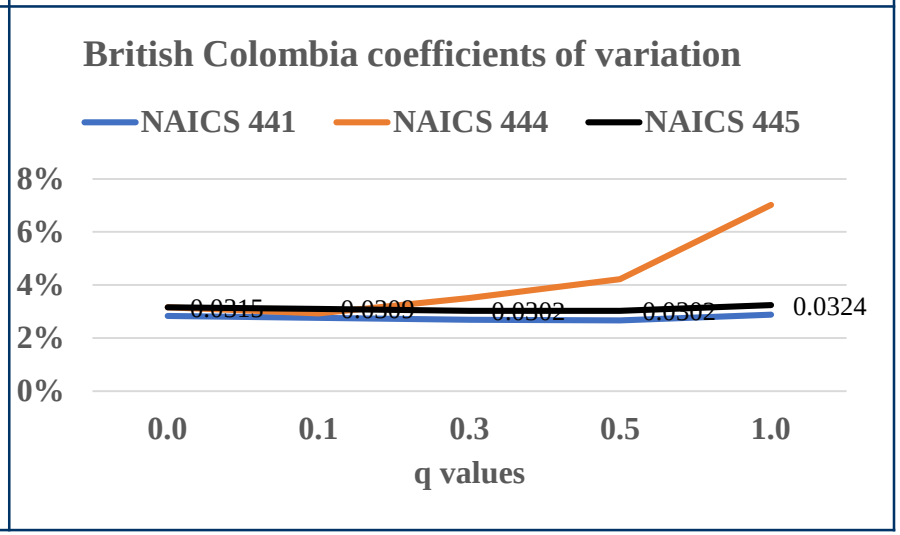


Results for *n*-scenario

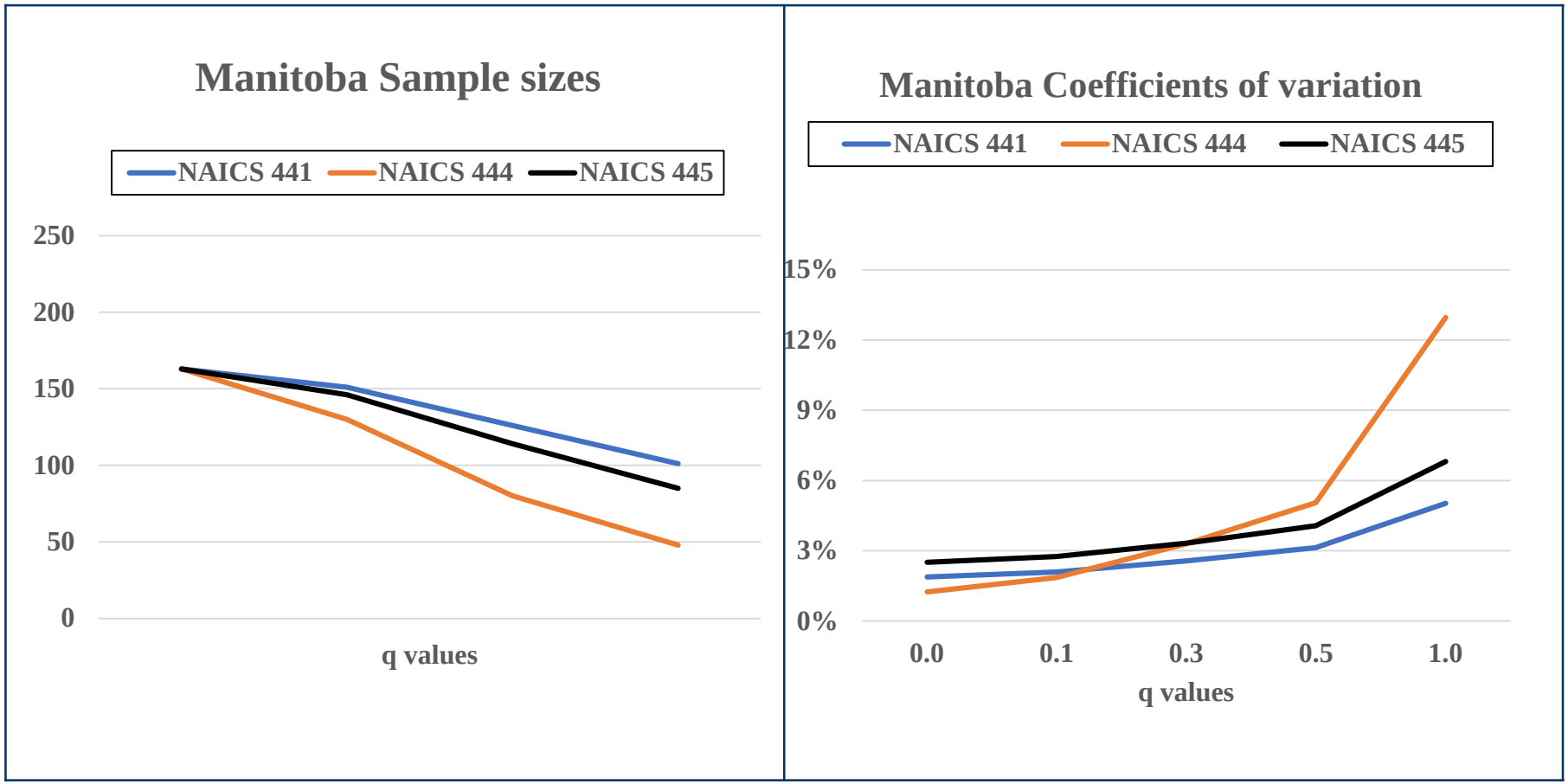








Results for *n*-scenario



Summary *n-scenario*

1. q = 0.1-0.3 appears optimal, as it balances sample size and coefficient of variation.
2. High q (≥0.5) leads to deterioration in CVs, especially for smaller provinces.
3. Large provinces tolerate higher q values without major loss in quality.
4. Comparable to *c*-scenario when q=0.1
5. Advantage over *c*-scenario: direct computation of sample sizes given n

3. Sampling from the Business Register

BR continually changing

- Need to record the status of statistical units targeted for sampling

Sample control file (SCF)

- central operational file used to manage and monitor the sample selection and estimation processes for a survey.
- Provides a structured record of all sampled units (e.g., businesses), their design information, and their evolving survey status.

3. Sampling from the Business Register

Sample control file (SCF)

- Each survey has its own SCF
- History of snapshots of the BR
- All statistical units in the scope of a survey are identified on the SCF
- SCF updated with each survey occasion to identify births and deaths of statistical units on the BR
- From survey occasion to survey occasion
  - Deaths are not deleted
  - Births are added

3. Sampling from the Business Register

Sample control file (SCF)

Typical components

- Survey ID
- Identification of statistical units
- Stratification: Geography, industry, size code
- Frame: Birth and death dates
- Sample selection: Inclusion flag, selection method, Initial weight ($1/\pi$), rotated in or out

SCF enables unbiased estimation for domains of interest

4. Removal of dead units on SCF

Why remove dead units?

- Dead units do not have impact on estimate
- However, variance of estimate gets larger over time if dead units are not removed

4. Removal of dead units on SCF

- Stratum is defined by the survey: geography, industry and size (employment / revenue)
- Basic idea: remove in-sample and out-of-sample dead units from latest Sample Control File
- Removal takes place at fixed dates: every 6 months

4. Removal of dead units on SCF

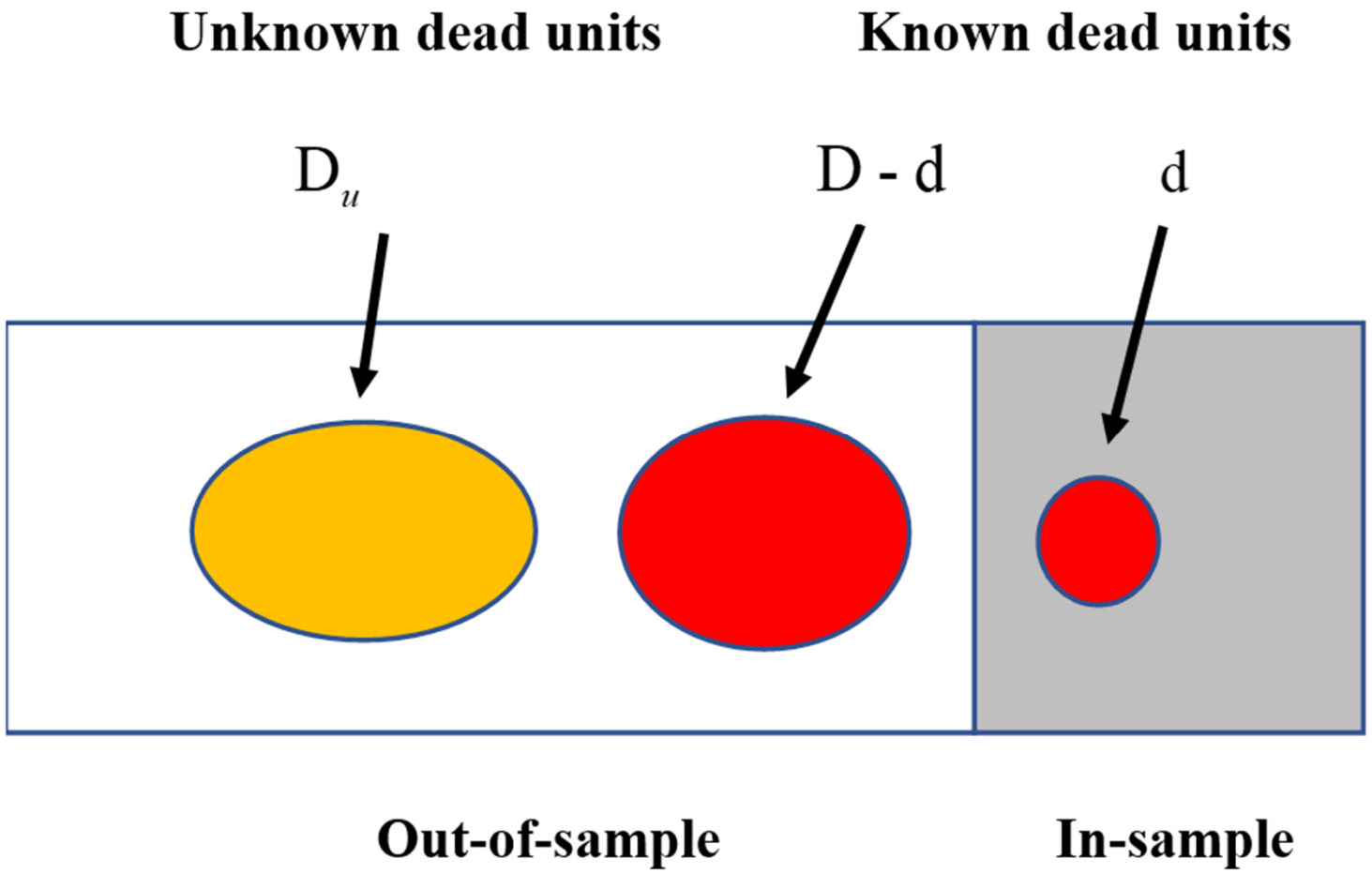
Problem:

How many dead units should be retained in sample and out-of-sample to represent the *unknown* number of out-of-sample dead units?

Solution

Remove a number in-sample *known* dead units, as well as a number of out-of-sample *known* dead units, so that the number of remaining in-sample dead units represents the out-of-sample *unknown* deaths.

4. Removal of dead units on SCF



4. Removal of dead units on SCF

Notation (stratum level)

N : population units

n : units in sample *s*

d : known deaths in sample

D-d : known deaths out-of-sample

4. Removal of dead units on SCF

w d	estimated number of deaths in the population
(w-1) d	estimated number of out-of-sample deaths
(w -1) d - (D – d)	estimated number of unknown dead units outside the sample

If (w -1) d - (D - d) ≤ 0: do nothing

If (w -1) d - (D - d) > 0: remove in and out-of-sample deaths

4. Removal of dead units on SCF

x : Number of dead units left in the sample:

w x: Estimated number of dead units in the population.

(w -1) x : Estimated out-of-sample deaths

Note that (w -1) d - (D-d) is equal to (w -1) x

Solving for x , we obtain

$$x = \mathrm{d} - \frac{\mathrm{D} - \mathrm{d}}{\mathrm{w} - 1}$$

4. Removal of dead units on SCF

1. Number of dead units removed from sample

$$d^* = \frac{\mathrm{D} - \mathrm{d}}{\mathrm{w} - 1}$$

2. Estimated number of deaths to be removed from the population

$$\mathrm{D}^* = \mathrm{w}\ \mathrm{d}^*$$

3. Remove $\mathrm{D}^* - \mathrm{d}^*$ units from the known D-d out-of-sample dead units

4. Removal of dead units on SCF

Outcomes for removing deaths

Condition	In-sample dead units removed	Out-of-sample dead units removed
$int(d^*) = 0$	None	None
$int(d^*) = d$	All	$int\ (D^* - d^*)$
$0 < int(d^*) < d$	$int(d^*)$	$int\ (D^* - d^*)$

5. Concluding Remarks

Focused on

1. The construction, structure, and ongoing maintenance of the business register
2. Methods for determining and allocating sample size
3. Sampling: Tracking is evolution, thereby ensuring unbiased (nearly) estimators of parameters of interest
4. A procedure for the elimination of dead units.