PRASC

Project for the Regional Advancement of Statistics in the Caribbean

Projet régional pour l'avancement de la statistique dans les Caraïbes

In **partnership** with Canada

Administrative Data use in Business Surveys

Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

Delivering insight through data for a better Canada

Alexander Reicker
March, 2022

Outline

- Administrative Data
  - Tax Data, Specialized data sources, Statistical Business Register
- Survey Steps
  - Frame Construction
  - Sampling
  - Collection
  - Edit and Imputation
  - Estimation
  - Confidentiality
- Other Considerations
  - Record Linkage

Statistical Infrastructure

- Statistical processes
- Classifications (industries, products, geography)
- Administrative data
- Statistical Registers (business, population, geography)
- Methodology
- Collection Instruments

Levels 1 and 2 of the Generic Statistical Business Process Model

Quality management / Metadata management							
1 Specify needs	**2 Design**	**3 Build**	**4 Collect**	**5 Process**	**6 Analyse**	**7 Disseminate**	**8 Evaluate**
1.1 Identify needs	**2.1** Design outputs	**3.1** Build collection instrument	**4.1** Create frame and select sample	**5.1** Integrate data	**6.1** Prepare draft outputs	**7.1** Update output systems	**8.1** Gather evaluation inputs
1.2 Consult and confirm needs	**2.2** Design variable descriptions	**3.2** Build or enhance process components	**4.2** Set up collection	**5.2** Classify and code	**6.2** Validate outputs	**7.2** Produce dissemination products	**8.2** Conduct evaluation
1.3 Establish output objectives	**2.3** Design collection	**3.3** Build or enhance dissemination components	**4.3** Run collection	**5.3** Review and validate	**6.3** Interpret and explain outputs	**7.3** Manage release of dissemination products	**8.3** Agree on action plans
1.4 Identify concepts	**2.4** Design frame and sample	**3.4** Configure workflows	**4.4** Finalise collection	**5.4** Edit and impute	**6.4** Apply disclosure control	**7.4** Promote dissemination products	
1.5 Check data availability	**2.5** Design processing and analysis	**3.5** Test production system		**5.5** Derive new variables and units	**6.5** Finalise outputs	**7.5** Manage user support	
1.6 Prepare business case	**2.6** Design production systems and workflow	**3.6** Test statistical business process		**5.6** Calculate weights			
		3.7 Finalise production systems		**5.7** Calculate aggregates			
				5.8 Finalise data files			

Statistical Classifications

International comparability – all countries have to follow the same standards to allow countries and governments to compare their performance.

- North America
  - North American Industry Coding System (NAICS)
  - North American Product Coding System (NAPCS)
- Caribbean
  - ISIC
- Geography

Administrative Data

- Data collected by other departments or organizations for their own purposes, used by Statistics Canada.
  - Tax and other data collected by Canada Revenue Agency
  - Data collected by provincial and municipal governments

- Generally Microdata related to a single business.

- Policy on the Use of Administrative Data Obtained under the *Statistics Act*.

Administrative Data

- Canada Revenue Agency (CRA) is the department responsible for most of the administrative data (via taxation):
    - **Business Number (BN) File**
    - **Business Tax Returns: T1 (unincorporated), T2 (incorporated)**
    - **Charitable and Other Returns: T1044 (non-profits), T3010 (charities), T5013 (partnerships)**
    - **Data on Goods and Services Tax / Harmonized Sales Tax**
    - **Data on personal income tax (T1)**
    - **Data on salary and wages paid to employees (T4 & PD7)**
    - **Data on some tax credit programs (e.g. R&D via T661)**
- There are also many other sources of admin data in Canada.

Enterprise Statistics Division

- Most CRA data at Statistics Canada is handled by ESD:
  - Acquisition
    - Development of Memoranda of Understanding
    - Receive the data
  - Processing
    - Basic clean up
    - Reformatting (example: monthly figures converted to annual)
  - Access and Uses
    - Need to know basis
    - System to apply, approve and review access
    - Track and control how the data are used

Administrative Data

Quality

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Interpretability
- Coherence

Tax Data

- Tax and other returns
- Data on revenues and expenses, personnel, tax credits, and other variables

- Used to populate the Business Register
- Used directly in survey processes

- Flash estimates and other data products

Other Admin Data Sources

Provincial and Territorial Data files
- Data collected from prov/terr statistical offices

Regulated Industries
- Mining Royalties (provincial)
- Telecommunications (federal)
- Energy

Industry Associations
- Pharmaceuticals
- Plastics

Frame Building

Data replacement

Edit and Imputation

Data Confrontation

Statistical Business Register

- In order to gather information on the economy, you need to have a current and complete frame of all units being observed (**businesses**, dwellings and geographies).
- Structure of the frame aligns with definitions and classifications.
  - A Statistical Business Register (SBR) is an inventory that essentially includes all businesses operating in your areas with some of their related information.
- **SBR should be the sampling frame for business surveys, and business surveys the 'frame' for economic statistics**
- But what's the 'frame' for SBR? **Admin data play a major role.**

Administrative data used to update SBR

- Coverage (updates on active/inactive enterprises)
- Classifications (industry, region, country of control, etc.)
- Business contact information, if agreements with partners.
  - Used for survey collection
- Main revenue and expenses variables for statistical use.
  - Size stratification, sampling allocation
  - Edit and Imputation
  - Estimation (e.g. calibration)

Statistical Business Register

View Update Utilities Help

Get: 102132 Go Statistical Search Contact Search

Funded by the Government of Canada
Financé par le gouvernement du Canada | Canada

Main Information

Statistical Number: 102129 Registrar ID: Social Security ID: TaxPayer ID: 70 Vat ID:

Business Status: Alive Legal Name: Gaetan Operating Name: Gaetan Bakery

Business Date: 05-Feb-2018 Fiscal Period: 01-Jan-2020 31-Dec-2020 Registration Date:

ISIC: 5629 SNA: 110001 Prospect Business Type: Corporation Country Of Control: JAM

Source	Name	Title	Phone Number	Email	Address	District	City	Country	Postal Code	Parish
IRD	Gaetan St-Louis	Mr.	1829214432							

Structure

- 102129 - Legal
  - 102130
  - 102131
  - 102132 - Legal

Business Information

Statistical Number: 102129 Enterprise Establishment Business Status: Alive Business Date: 05-Feb-2018

Registrar ID: Social Security ID: TaxPayer ID: 70 Vat ID:

Legal Name: Gaetan Operating Name: Gaetan Bakery

ISIC: 5629 Bakery café ISIC Effective Date: 05-Feb-2018

Legal Address:

35 Center St
Port Antonio
Port Antonio
St. George

Operating Address:

35 Center St
Port Antonio
Port Antonio
JAM

Business Size Information

Revenue:	-256789	2018	IRD
Sales:			
Wages:			
Number of Employees	4	2020	On-line
Male Employees:	2	2020	On-line
Female Employees:	2	2020	On-line

Legal Phone Number: 1829214432

Operating Phone Number: 1829214432

Website:

Comment:

Type here to search -3°C Mostly cloudy ENG CMS 12:49 PM 2022-03-16

Statistical Business Register

View Update Utilities Help

Get: 102132 Go Statistical Search Contact Search

Funded by the Government of Canada
Financé par le gouvernement du Canada | Canada

Main Information

Statistical Number: 102129 Registrar ID: Social Security ID: TaxPayer ID: 70 Vat ID:

Business Status: Alive Legal Name: Gaetan Operating Name: Gaetan Bakery

Business Date: 05-Feb-2018 Fiscal Period: 01-Jan-2020 31-Dec-2020 Registration Date:

ISIC: 5629 SNA: 110001 Prospect Business Type: Corporation Country Of Control: JAM

Source	Name	Title	Phone Number	Email	Address	District	City	Country	Postal Code	Parish
IRD	Gaetan St-Louis	Mr.	1829214432							

Structure

- 102129 - Legal
  - 102130
  - 102131
  - 102132 - Legal

Business Information

Statistical Number: 102130 Enterprise Establishment Business Status: Alive Business Date: 05-Feb-2018

Registrar ID: Social Security ID: TaxPayer ID: Vat ID:

Legal Name: Operating Name: Gaetan Bakery

ISIC: 5629 Bakery Café ISIC Effective Date: 05-Feb-2018

Legal Address:

Operating Address: 60 Center St, Port Antonio, Port Antonio, St George, JAM

Legal Phone Number:

Operating Phone Number: 1829214432

Website:

Comment:

Business Size Information

Revenue:	700125	2020	On-line
Sales:			
Wages:			
Number of Employees	2	2021	On-line
Male Employees:	1	2021	On-line
Female Employees:	1	2021	On-line

Type here to search -3°C Mostly cloudy ENG CMS 12:50 PM 2022-03-16

Business Survey Steps

Frame building

- **SBR, external lists**

Sampling

- **sample stratification, allocation, and selection**

Collection

- **contact information**

Edit and imputation

- **identify outliers, data replacement**

Estimation

- **calibration, small area estimation, macro adjustments**

Dissemination

- **confidentiality**

Frame building

In NSO, we can use SBR to get the most up-to-date information on:

- **Business numbers or identifiers**
- **Business activity status (coverage)**
- **Classifications (ISIC or region)**
- **Size variables (revenue or # of employee)**
- **Contact information (phone and e-mail)**
- **Statistical levels (enterprise or establishment)**

External lists coming from external agencies/departments

- Special business surveys
- Secondary activities

Coverage and Classification Errors

Risks:

- Target Population ≠ Sample Population
- Out of scope / Inactive units
- Stratification and allocation errors

Sampling

- Sampling population in business surveys
  - Capitalize on the skewed nature of the business population
    - 90% of activity ➔ 10% of the population
  - Consider the publication domains
    - Typically industry and geography

- Stratification groups similar records together to improve sampling efficiency
  - Inaccurate / outdated information quickly leads to strata not containing similar units

    ➔ loss of survey efficiency and loss of accuracy

Sampling

- Typical strata (could differ depending on the type of survey):
  - Must take - Subject Matter specified units that must be selected
    - Often have unique characteristics, or are very large
  - Take all - Largest units, most important to an industry should always be included
  - Take some - Mid-sized to large units, sampled based on thresholds
  - Take few – Mid-sized to small units,
  - Take none - Smallest units, least important, could be excluded
    - Often a large number of units accounting for very little activity
    - Macro adjustment as part of the estimation process
- *May have an extra layer of stratification to ensure sufficient responses for quality estimates for geographic areas*

Administrative Data Hierarchy

Multiple Sources for same concept (e.g. Revenue):

- Profiled Revenue
  - ➢Not available for all units
- Different tax years
  - ➢Most Recent
- Business Return vs Sales Tax Return
  - ➢Total Revenue model based on Sales Tax

Sampling Stratification

Size stratification: Meet variance (quality) targets with a smaller sample.

Size measure should be strongly correlated with variables of interest.

Unit ID	Selected	Weight	True Revenue	Estimation Contribution
1	Yes	5/3	10 000	16 667
2	Yes	5/3	20 000	33 333
3	Yes	5/3	500	833
4	No	0	200	
5	No	0	100	
Total		5	30 800	50 833

Sampling Stratification

Size stratification: Meet variance (quality) targets with a smaller sample.

Size measure should be strongly correlated with variables of interest.

Unit ID	SBR Size	Selected	Weight	True Revenue	Estimation Contribution
1	9 000	Yes	2/2 = 1	10 000	10 000
2	15 000	Yes	2/2 = 1	20 000	20 000
3	800	Yes	3/1 = 3	500	1 500
4	400	No	0	200	
5	400	No	0	100	
Total	25 600		5	30 800	31 500

Size Stratification Variable Errors

Risks:

- Stratification and allocation errors
- High variance
- Take-None population

Collection and non-response

- Having the most up-to-date contact information is important to get a good response rate and get the best quality we can out of our selected sample:
  - **Unreachable units could create bias if there is a non-response pattern (e.g. small units don't have e-mails).**
  - **Reach the right entity, respondents, level (avoid response bias)**
  - **Adding collection edits ("Based on our information, you reported $X revenue last year. Please confirm your answer.").**

Collection and non-response

- Having the most up-to-date contact information is important to get a good response rate and get the best quality we can out of our selected sample:
  - **Help non-response (or partial) follow-up, e.g. questionnaire sent by e-mail, but follow via phone call**
  - **Could use SBR revenues to flag the cases that would require follow up for survey nonresponse.**
    - **Could then impute missing values with SBR values.**
  - **Not spending time surveying out-of-scope and out-of-business units**

Collection and non-response

"Short form" questionnaires

- Tax replacement to complete the financial portion of the questionnaire
- Only valid where the tax unit is equivalent to the collection unit

Pseudo-census approach

- Use imputation (based on tax) instead of weights for the non-sampled part of the population
- No change to sample stratification, allocation or selection

Contact Information Errors

Risks:

- Target Population ≠ Sample Population
- Out of scope / Inactive units

Edit and Imputation

Editing: identify errors or inconsistencies in collected values

- (Total Operating Expenses) + (Interest Expenses) = (Total Expenses)

Imputation: complete missing data

- Partial: some fields in the questionnaire not filled out
- Total: the whole questionnaire is empty

Edit and Imputation

In business surveys we normally impute for non-response (partial or complete), mainly because

- we have a lot of auxiliary information (SBR, admin data, historical values) to develop proper imputation models
- Reweighting for total non-response could introduce high bias as we tend to make follow-ups on the biggest companies that did not answer

Edit and Imputation

- Direct imputation of missing values
  - Concepts must match exactly
  - Admin data itself may be imputed
  - Will work for all units in the population
- Trend calculation
- Nearest Neighbour criteria for donor imputation
- Geography, Industry and Size definition for imputation classes

Administrative data is often considered to be of equivalent quality to reported data.

Estimation

- Estimation domains should align with sampling strata (e.g. regions, ISIC). When sample design is well planned, most estimates should be of good quality (depending on the targeted quality)
- Admin data can be used in:
  - Outlier Detection: compare reported revenue to admin data sources. Large differences indicate that the weight should be adjusted.
  - Calibration: adjust weights based on available admin data totals to reduce volatility of estimates (e.g. bad random sample) in exchange for a (small) bias.
  - Take-None Modelling: Ratio model based on relationship between admin data and survey variables for the TN part of the population.
  - :Small Area Estimation: When a sampling stratum covers many estimation domains, admin data can be used to build a statistical model.

Calibration

- Adjust the weights to better reflect other known information about the population
- Need auxiliary information for all of the units on the frame
- Reduces the variance, but could introduce (asymptotic) bias

Calibration

Auxiliary information:

- Available for all of the units on the frame
- Well correlated with the survey variables
- Admin data or from a census
    - Number of businesses on SBR
    - Total tax revenues
    - Number of farm related acres

Calibration modifies the weights optimally so that the estimated total using the weights equals the known total.

Take-None Modelling

Take-None units are the smallest units in the population

- Not sampled
- Contribute < 10% of total revenue
- Most similar to small and medium Take-Some Units

Take-None Modelling

Take-None model

- Administrative data (revenue and other variables) available for all Take-None units
- Simple ratio model based on small and medium Take-Some units
- Aggregated to the domain level to be added during estimation
- Admin and survey variables must be correlated

Confidentiality

- Identify sensitive cells
  - Avoid identifying individual businesses
  - Dominance within a cell
  - Enterprise level

- Protect the confidential data collected from disclosure
  - Cell suppression
  - Secondary Suppression
  - Random Tabular Adjustment

Confidentiality

Enterprise ID	Establishment ID	Value
1	1	1 000
2	2	1 000
3	3	2 000
3	4	2 000
3	5	3 000
3	6	3 000
3	7	8 000
	Total	20 000

Enterprise ID	Value
1	1 000
2	1 000
3	18 000
Total	20 000

Dominance!

Record Linkage

- Direct Linkage Key
  - Unique Identifier (Business Number)

- Indirect Linkage Variables
  - Business Name, Address

Conclusion

Primary uses of Administrative Data:

- Frame Construction
- Stratification
- Contact information
- Take-None modelling
- Calibration
- Confidentiality

Secondary uses of Administrative Data:

- Collection Follow-up
- Edit and Imputation

You can contact the PRASC team at:

prasc@statcan.gc.ca

Presenter: Alexander Reicker
alexander.reicker@statcan.gc.ca