The Impact of Processing Administrative Sources on the Quality of Statistical Outputs

Martin Beaulieu,
Chief - Quality and Data Ethics Secretariat, Statistics Canada
Global Seminar on Administrative and Other Data Sources
November 14, 2023

Delivering insight through data for a better Canada

Outline

- Context
  - Quality of inputs – Canadian context
  - Quality of process
- Statistics Canada use-case: Canadian Housing Statistics Program (CHSP)
  - Administrative sources in Canada
  - Processing required to integrate sources and derive estimates for CHSP
  - Quality Indicators

Administrative Sources in Canada

- Three levels of government:
  - Federal (national)
  - 10 provinces and 3 territories
  - Over 5000 municipalities (local, established under provincial/territorial authority
- Level of jurisdiction changes on the subject:
  - For example, education is a provincial jurisdiction, defense is federal, transportation is shared

Quality of Process

- Quality of input is essential for any statistical program
- Administrative sources are usually of good quality
- Integrating multiple sources from different jurisdictions can be complex:
  - Concepts definitions may vary
  - Different code sets may be applied for a same variable
  - Processing to standardize to federal definitions
    - For example, Geography has to be standardized to the Census definitions.
- Each step of the process to produce final estimates can potentially introduce errors

Concrete Example

- Canadian Housing Statistics Program (CHSP)
  - Administrative census of residential properties and their owners.
- Disseminates statistical information about the residential housing sector at the municipal level
  - Number and type of properties
  - Assessment value
  - Total living area
  - Property use (owner-occupied or not)
  - Residency ownership (resident or non-resident)
- Integration of multiple sources of administrative data
  - Provincial and territorial land registries
  - Tax data of property owners (city-level)
  - Business Register
  - Census of Population
  - Longitudinal Immigration Database

Processing required

Geography/domains, estimates	Individual Quality indicator
Geography	Geocoding rate, Geocoding confidence score
Property Type	Coding Rate
Property Use	Linkage Error Rate
Residency Ownership	Linkage Error Rate
Period of Construction	Coding Rate
Total Living Area	Reporting Rate, Inclusion Rate
Residency Participation	Linkage Error Rate
Property Assessment Value	Reporting Rate

Quality Indicators

- Frameworks such as the Total Survey Error Framework and its extensions (Zhang, 2012) and (Reid et al., 2017) as well as the UN-NQAF Manual for official statistics and Statistics Canada's Quality Guidelines provide guidance on quality indicators to derive at different steps of the data processing
- How can these indicators be used to communicate quality in a way that is clear for the users?
  - The CHSP had to objective to have one single indicator with each estimate to inform users on the quality of this estimate
  - With multiple indicators, it becomes a multi-dimensional problem to solve

Quality Indicators

- Clustering techniques were used to address this challenge
- Domains for which estimates have a similar overall level of quality are grouped together
  - The steps to develop the Composite Quality Indicators (CQI) are as follow:
    1. Standardization of quality indicators
    2. Weighting of quality indicators (using ANOVA results)
       - Under the principle that a classification error in a domain variable has more impact on the quality of an estimate if the domain variable is strongly associated with the estimated parameter
    3. K-means clustering
- Once clustering has been completed, a global score for each cluster is calculated to draw a profile of each cluster to better understand how they differ from each other in terms of the quality indicator values.
- The cluster with the highest global score is assigned the value A, the cluster with the second highest global score is assigned the value B, etc.
- Data Visualization can help interpret the results

Quality Indicators

- CQI conclusion and limitations:
  - Clustering is simple and fast way to summarize a large number of quality indicators in a categorical quality rating.
  - The purpose of the CQI is to provide indication of the overall quality of an estimate to enable its use, but will not enable inference
  - Ability to interpret the results is key. Since the CQI does not provide an absolute value of quality, is important to explain that the CQI assigns a level of quality to a group of domains relative to the other groups.
  - Despite the relative complexity of interpretation of the CQI compared to indicators based on CVs, this method provides a good indication of the overall quality of an estimate and gives a better view of the global picture of the quality of data processing steps.
- For more details on the CQI method: Development of a composite quality indicator for statistical products derived from administrative sources (statcan.gc.ca)

Thank you/ Merci

Pour de plus amples renseignements, veuillez contacter:

For more information, please contact:

martin-j.beaulieu@statcan.gc.ca