Creation of a Composite Quality Indicator for Estimates Based on Administrative Data Using Clustering

Roxanne Gagnon, Martin Beaulieu

Statistics Canada, Canada

Delivering insight through data for a better Canada

1. Context

- Statistics Canada, like many national statistical agencies, has begun the transition to integrated models where probability surveys are no longer used alone, but in combination with administrative or alternative data sources (e.g. big data, remote sensing, web scraping).

- Data integration provides a unique opportunity to improve the quality of official statistics in terms of timeliness and potentially accuracy, and produce more disaggregated statistics.

- However, measuring and reporting accuracy is a significant challenge, as the methods and terminology used by national statistical agencies are still largely rooted in sampling theory.

- The goal of the composite quality indicator (CQI) is to communicate the accuracy of the estimates to data users in the context of a statistical program based entirely on administrative data sources.

2. Reporting data quality at Statistics Canada

- Principle 7 of Statistics Canada's Quality Assurance Framework [1] and the Policy on Informing Users of Data Quality and Methodology [2]: **Users must be informed of data quality to be able to assess fitness for use.**

- Metadata and technical reports are important, but it is also recommended to use a letter grade attached to each estimate based on the sampling variance. For example, one statistical program could use:

Symbol	Label	Coefficient of variation (%)
A	Excellent	0 – 4.99
B	Very good	5 – 9.99
C	Good	10 – 14.99
D	Acceptable	15 – 24.99
E	Use with caution	25 – 34.99
F	Too unreliable to be published	35 and over

2. Reporting data quality at Statistics Canada

- The same approach cannot be used for estimates based on administrative data
  - There is no sampling error
  - Like for surveys, most types of non-sampling error are difficult to measure

- But other information on accuracy are available internally since Statistics Canada recommends in its Quality Guidelines [3] to use quality indicators (QIs) to control the quality of inputs and processes, for example:

Quality of inputs	Quality of processes
Reported rates of variables	Coding rates of variables
Coverage rates of source files	Record linkage error rates

- How can these QIs be used to communicate quality in a way that is clear for users?
- The proposed CQI approach is to use clustering to obtain a single categoric value for each estimates.

3. Canadian Housing Statistics Program (CHSP)

- 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
  - Business Register
  - Census of Population
  - Longitudinal Immigration Database

4. Selecting quality indicators for each survey variable

DOMAIN VARIABLES				VARIABLES OF INTEREST		
Geography		Type of properties, and period of construction	Property use, and residency ownership	Assessment value	Total living area	
Geocoding rate	Quality score of geocoding	Coding rates	Record linkage quality	Reported rate	Reported rate	Inclusion rate
			Precision (1 - false discovery rate)	Recall (1 - false negative rate)		

4. Selecting quality indicators for each survey variable

Some observations about the QIs:

- QIs are measured at domain level so they are rates or averages (e.g. geocoding quality score).
- They vary in the same direction: the higher the value of the QI, the better the quality is.
- They are bounded between 0 and 1, and the distributions are heavily skewed toward 1.
- In the CHSP example, they were found to be mostly uncorrelated.
- Variance of QIs must be greater than 0 to be used in clustering.

Prior to clustering, each QI is standardized and multiplied by a factor (importance weight):

- The goal is to give more importance to some QIs that are deemed to be more important for a given estimate.
- Importance weights are first calculated from ANOVA models and then validated by subject matter experts.

5. Combining the quality indicators: The importance weights

- Analysis of variance (ANOVA) models with fixed effects, using microdata (properties).

Fixed effects (Domain variables)	Effect size (%) on each continuous variables of interest		
	Assessment value	Total living area	Assessment value by square foot
Census subdivision	**75.4**	**24.4**	**95.7**
Property type	**20.8**	**60.1**	0.7
Period of construction	3.2	**14.0**	3.2
Ownership type	0.6	0.5	0.0
Property use	0.0	1.0	0.4
Residency ownership	0.0	0.0	0.0
Residency participation	0.0	0.0	0.0

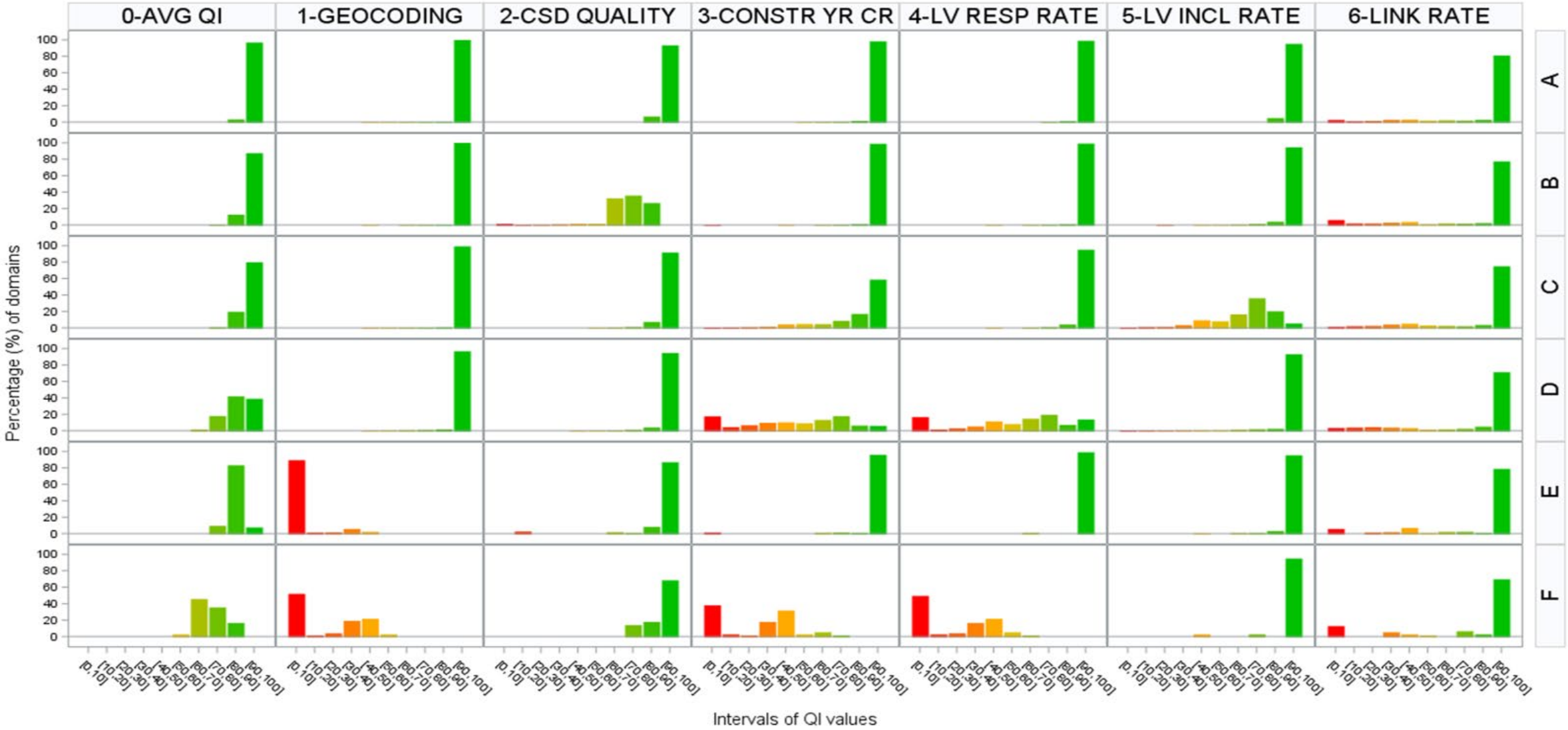
5. Combining the quality indicators: The order of the clusters

- K-means clustering of the weighted QIs was used to group domains that are similar in terms of the different QIs, but the resulting clusters are unordered.

- We want to order them to facilitate the interpretation:

1. A global score (the weighted average) is calculated as a measure of the average quality in each cluster.
2. Clusters are then ordered based on the global score:
   A is the best cluster
   B is the second best cluster
   Etc.
3. The profile of the clusters are visualized.

Distribution of Quality Indicators by value of the Composite Quality Indicator for Table 54 - Total living area, average and median



6. Interpreting the composite quality indicator values

Symbol	Label	Quality of components
A	Excellent	All QI components of the CQI are deemed excellent.
B	Very good	The CQI is deemed very good due to the very good geolocation quality and high levels of quality of the other components.
C	Good	The CQI is deemed good due to the good geolocation quality and high levels of quality of the other components.
D	Acceptable	The CQI is deemed acceptable due to the low level of quality for the geolocation or the period of construction coding while all remaining quality indicator components have a high level of quality.
E	Use with caution	The CQI has a quality that prompts caution for the use of the estimate.
F	Too unreliable to be published	-

7. Strengths and challenges

Strengths

- **Simple.** Combining QIs with statistical methods that are well known and implemented in the majority of statistical software applications.
- **Fast.** K-means clustering is faster than other unsupervised learning algorithms for large data sets.
  - In CHSP: 100,000-1,000,000 records
- QIs can be simple.
- Importance weights are obtained in an objective manner.
- Groups are created automatically.

Challenges

- **Data-driven.** Quality levels obtained are relative and not absolute.
- Multiple clustering models are needed, which limit comparability between tables, estimated parameters.
- Documentation must be provided to users to guide them with the interpretation of the CQI values.

8. Conclusion and future works

- Until a year ago, no statistical program released quality indicators for estimates exclusively based on administrative data.
- The Canadian Housing Statistical Program is the first statistical program to use the combination of data processing QIs to inform users about quality.
  - The CQI was included for the first time in a September 2021 release. The documentation of the approach for users was published last January. [4]
  - 8 tables of results for properties, owners and buyers now incorporate CQIs.
- This is a significant advance in assessing the quality of estimates produced from administrative data sources.
- Research work is continuing to improve the QI components of the CQI and to develop new ones.

Merci ! — Thank you!

Pour de plus amples renseignements, veuillez contacter /
For more information, please contact
roxanne.gagnon@statcan.gc.ca

The content of this presentation represents the position of the author and may not necessarily represent that of Statistics Canada.

References

[1] Statistics Canada (2017). *Statistics Canada's Quality Assurance Framework*, Catalogue no. 12-586-X. https://www150.statcan.gc.ca/n1/en/catalogue/12-586-X

[2] Statistics Canada (2000). *Policy on Informing Users of Data Quality and Methodology*. https://www.statcan.gc.ca/en/about/policy/info-user

[3] Statistics Canada (2019). *Statistics Canada Quality Guidelines*, Sixth edition, Catalogue no. 12-539-X. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm

[4] Statistics Canada (2022). *Development of a composite quality indicator for statistical products derived from administrative sources,* Catalogue no. 46-28-0001. https://www150.statcan.gc.ca/n1/en/catalogue/46280001202200100001