Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society’s Emerging Needs

Creation of a composite quality indicator for administrative data-based estimates using clustering

by Roxanne Gagnon, Martin Beaulieu, Danielle Lebrasseur,
Wei Qian and Anthony Yeung

Release date: October 22, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Creation of a composite quality indicator for administrative data-based estimates using clustering

Roxanne Gagnon, Martin Beaulieu, Danielle Lebrasseur, Wei Qian and Anthony Yeung[1]

Summary

National statistical agencies such as Statistics Canada have a responsibility to convey the quality of statistical information to users. The methods traditionally used to do this are based on measures of sampling error. As a result, they are not adapted to the estimates produced using administrative data, for which the main sources of error are not due to sampling. A more suitable approach to reporting the quality of estimates presented in a multidimensional table is described in this paper. Quality indicators were derived for various post-acquisition processing steps, such as linkage, geocoding and imputation, by estimation domain. A clustering algorithm was then used to combine domains with similar quality levels for a given estimate. Ratings to inform users of the relative quality of estimates across domains were assigned to the groups created. This indicator, called the composite quality indicator (CQI), was developed and experimented with in the Canadian Housing Statistics Program (CHSP), which aims to produce official statistics on the residential housing sector in Canada using multiple administrative data sources.

Keywords: Unsupervised machine learning, quality assurance, administrative data, data integration, clustering.

1. Introduction

Users must be informed about the quality of the data so they can determine whether the statistical information is appropriate for their intended use. This is a guiding principle of quality assurance at Statistics Canada, as stated in the Quality Assurance Framework (Statistics Canada 2017) and the Policy on Informing Users of Data Quality and Methodology (Statistics Canada 2000). There are a number of ways that this requirement is met, such as the publication of technical reports and metadata for surveys that help users judge all dimensions of quality (relevance, timeliness, accuracy, coherence, interpretability and accessibility). It is also recommended that a measure of the accuracy of each estimate be reported. For sample surveys, a measure of sampling error is published, such as a confidence interval, a coefficient of variation, or a rating from A to F derived from the coefficient of variation.

In recent years, national statistical agencies have begun transitioning to integrated models where probabilistic surveys are no longer used alone, and are instead used in combination with other administrative or alternative data sources (e.g., big data, remote sensing, web scrapping). These new data sources can be used to supplement sample surveys, be completed by sample surveys, or even replace sample surveys altogether. Data integration affords a unique opportunity to improve the quality of official statistics in terms of timeliness and potentially accuracy. However, measuring and reporting accuracy is a significant challenge in this context, as the methods and terminology used by national statistical agencies are still largely rooted in sampling theory.

Work has been done at other organizations to adapt the total survey error paradigm to integrated data (Zhang 2012), administrative data (Reid et al. 2017) and big data (Amaya et al. 2020). These studies compiled an exhausted list of potential sources of error and adapted existing terminology to these new data sources. However, they offer few concrete recommendations on how to report these errors quantitatively and in terms of their impact on the estimates

[1] Roxanne Gagnon, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (roxanne.gagnon@statcan.gc.ca); Martin Beaulieu, Statistics Canada (martin-j.beaulieu@statcan.gc.ca); Danielle Lebrasseur, Statistics Canada, (danielle.lebrasseur@statcan.gc.ca); Wei Qian, Statistics Canada, (wei.qian@statcan.gc.ca); Anthony Yeung, Statistics Canada, (anthony.yeung@statcan.gc.ca).

The content of this paper reflects the positions of the authors, not necessarily that of Statistics Canada.

in order to inform users. At Statistics Canada, the best practices recommended in the Quality Guidelines (Statistics Canada 2019) involve deriving indicators that reflect the quality of the statistical information production processes to judge their performance and their potential impact on estimates. Individually, these indicators are simple to understand and measure. When considered as a whole, it can be difficult to draw a clear enough picture to convey to users. The composite quality indicator (CQI) approach proposes to use unsupervised machine learning to combine the indicators into a single value, similar to that commonly used in sample surveys.

2. Canadian Housing Statistics Program

The CQI was developed to report on the accuracy of estimates from the Canadian Housing Statistics Program (CHSP). This statistical program uses administrative data sources exclusively to produce statistics in order to monitor and analyze the Canadian residential housing market. These data have varying levels of quality when they are acquired and the various steps to process and produce final estimates can potentially introduce errors. Three output tables (Statistics Canada 2021) with multiple estimates were targeted to develop the CQI. The estimation domains are defined based on many categorical variables and a geographic variable presented at several levels, the most detailed being the census subdivision (CSD).[2]

3. Creating the composite quality indicator

3.1 Selecting quality indicators

The first step was to identify the quality indicators of the processes used to produce the variables involved in constructing the different tables. In the CHSP example, the accuracy of the qualitative variables was represented by the geocoding rate, the average geocoding confidence score, coding rates, and record linkage error rates, i.e., the false discovery rate (FDR) and false negative rate (FNR). The accuracy of the quantitative variables was represented by reporting and inclusion rates. These indicators are summarized in Table 3.1-1. They are estimated at the domain level.

Table 3.1-1
Selected quality indicators

Variables[3]	Quality indicator
Geography	Geocoding rate, Average geocoding confidence score
Construction period	Coding rate
Total living area	Reporting rate, Inclusion rate
Owner type	Coding rate
Property type	Coding rate
Property use	Accuracy (1 - FDR), Recall (1 - FNR)
Property assessment value	Reporting rate

3.2 Clustering

The quality indicators were combined using k-means clustering. This algorithm was used to combine estimation domains with similar levels for the various indicators. It uses a distance measure that represents the dissimilarity between estimation domains, considering all quality indicators simultaneously. A diagnostic chart was used to select the optimal number of clusters. Other model specifications, i.e., the choice of quality indicators to include as well as the weights assigned to each indicator, were determined based on the type of estimation and the variable of interest.

[2] 2016 Census of Population Dictionary: https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/geo012-eng.cfm
[3] Integrated metadatabase: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvVariableList&Id=1322247

3.3 Weighting

The quality indicators were first standardized to remove the effect of scale, then weighted to control for the importance given to each indicator during clustering. Two weighting methods were used: one for counts and percentages, and one for continuous variables of interest.

For counts and percentages, equal weights were assigned to each domain variable, then the weight of each variable was distributed equally among the quality indicators related to that variable. Let us suppose a table constructed from D domain variables designated as X_d for d = 1, …, D. The accuracy of each variable X_d is represented by the quality indicators $IQ_{X_{dj}}$ for j = 1, …, J_{X_d}. The weight $w_{X_{dj}}$ of $IQ_{X_{dj}}$ is given by:

$$w_{X_{dj}} = \frac{1}{D \cdot J_{X_d}}$$

For totals, means and medians of continuous variables of interest, the weights of the domain variables were distributed proportionally according to the strength of the relationship between the domain variable and the variable of interest. Let us suppose a results table constructed from a continuous variable of interest Y and D domain variables. The weight of the quality indicators IQ_{Y_j} for j = 1, …, J_Y is given by w_{Y_j}:

$$w_{Y_j} = \frac{1}{(D+1) \cdot J_Y}$$

The weights of the quality indicators $IQ_{X_{dj}}$ are given by:

$$w_{X_{dj}} = \frac{D}{(D+1)} \cdot \frac{\eta_{X_d}}{\sum_{\mathrm{d}=1}^{\mathrm{D}} \eta_{X_d}} \cdot \frac{1}{J_{X_d}}$$

where η_{X_d} is the effect of the size of the variable X_d in the ANOVA model of Y as a function of the domain variables. Table 3.3-1 shows the results of the ANOVA models run on the microdata (properties) for the domain and interest variables in the selected CHSP tables.

Table 3.3-1
Results of the ANOVA models

Effect (X_d)	Effect size on variable of interest Y (%)		
	Property assessment value	Total living area	Property assessment value per sq. ft.
Construction period	1.1	5.7	0.9
CSD	26.1	9.9	26.7
Owner type	0.2	0.2	0.0
Property type	7.2	24.4	0.2
Property use	0.0	0.4	0.1
Residual	65.3	59.4	72.0

As can be seen in Table 3.3-1, CSD was the domain variable that accounted for most of the observed variance (26.1%) in the property assessment value. Therefore, the quality indicators related to CSD were most important in creating the CQI for the estimates of the mean and median of the property assessment value. For the mean and median of total living area, it was the quality indicators related to property type that were expected to be the most important since this variable accounted for most of the variance (24.4%) in total living area. The domain variables as a whole represented only a portion of the variance in the three variables of interest. The size of the effects could therefore be biased by omitting some variables. However, since the objective was to produce a measure of the relative importance of the domain variables in the estimation and not to produce an explanatory or predictive model, adding other variables or interactions to the model was not useful.

3.4 Order of the groups

The groups produced by clustering were not in order. They were therefore assigned an order to facilitate interpretation. This order was based on an overall score calculated for each group k as follows:

$$\mathrm{Score}_k = \frac{\sum_{i=1}^{M} \mathbb{1}_{ik} \cdot \overline{\mathrm{IQ}}_i}{\sum_{i=1}^{M} \mathbb{1}_{ik}}$$

where M is the total number of estimation domains, $\mathbb{1}_{ik}$ is a function that takes the value 1 if estimation domain i has been assigned to group k, otherwise 0, and $\overline{\mathrm{IQ}}_i$ is the weighted average of the quality indicators:

$$\overline{\mathrm{IQ}}_i = \sum_{\mathrm{j}=1}^{J_Y} w_{Y_j} \cdot IQ_{Y_{ji}} + \sum_{d=1}^{D} \sum_{\mathrm{j}=1}^{J_{X_d}} w_{X_{dj}} \cdot IQ_{X_{dji}}$$

The group with the highest overall score was designated as group A, the group with the second highest score was designated as group B, etc.

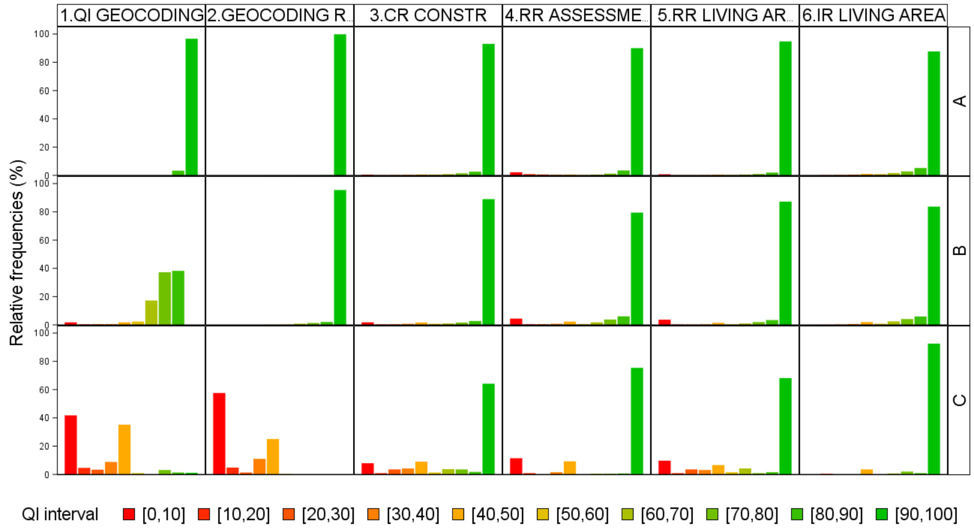
3.5 Visualizing group profiles

To properly interpret the CQI values, it is imperative to visualize the results to understand what distinguishes the groups from each other. Figure 3.5-1 on the next page is an example of a visualization developed for study purposes. Each cell in the grid shows the relative frequency distribution of a given quality indicator among the columns and for a given group among the rows.

Figure 3.5-1 shows that groups A and B are different by the average geocoding confidence score (first column). It is almost always above 90% for group A (first row) and between 60% and 90% for group B (second row). Group C (third row) differs from the other two groups in both the average geocoding confidence score (first column) and the geocoding rate (second column), which are both below 50%.

In the context of the CHSP, multiple clustering models were required to describe the accuracy of the estimates, as the selected result tables included multiple types of estimates and more than one variable of interest. A data visualization such as that shown in Figure 3.5-1 was performed for each clustering model required. The final scores assigned to each group in each model were determined after comparing the graphs of the different models.

Figure 3.5-1
Example of data visualization — Relative frequencies of quality indicators in each group



Acronyms: CR = Coding rate, IR = Inclusion rate, RR = Reporting rate

3.6 Final label assignment and interpretation

Table 3.6-1 presents the information provided to users to help them judge the suitability of CHSP estimates for use. Each CQI value is associated with a standard label. A description of the quality indicators, referred to as the components of the CQI, is provided for each group.

Table 3.6-1
Composite quality indicator values, label, and description of components

CQI value	Label	Description of components
A	Excellent	All components of the CQI are rated as excellent.
B	Very good	The CQI is rated as very good due to the very good quality of geocoding and the excellent quality of the other components.
C	Good	The CQI is rated as good due to the good quality of geocoding and the excellent quality of the other components.
D	Acceptable	The CQI is rated as acceptable due to the low quality level of geocoding or construction period coding, while the quality level of the other components is excellent.
E	Use with caution	The CQI indicates a level of quality that calls for caution in use.
F	Too unreliable to publish	-

The use of clustering to create the CQI implies that the values obtained are considered relative, not absolute, i.e., the accuracy of an estimate in one domain is not evaluated against a pre-established standard, but rather in comparison to estimates in other domains in the table. It would be difficult to establish standards for different programs, tables, types of estimates, or variables of interest that would be both consistent and relevant, since the quality indicators and weights are not always the same across estimates. Despite this, clustering is the most objective way to create quality groups and avoid using arbitrary rules.

4. Limitations

The limitations of the CQI relate primarily to the quality indicators used for clustering. In the context of the CHSP, some indicators measure the proportion of administrative units for which a processing step was completed, but not necessarily the reliability of that step. For example, a value may be coded, but the probability that the coded value is accurate is unknown. Furthermore, measuring record linkage error rates requires manual verification of a sample. It would be time-consuming and expensive to select and verify a sample large enough to obtain them at the CSD level. In the context of CHSP, linkage error rates were only available at higher geographic levels, which limited the effectiveness of these quality indicators in clustering models.

5. Conclusion

Clustering is a simple, fast and efficient approach to combine a series of quality indicators based on data processing steps such as geocoding, record linkage and imputation. The quality indicators can be weighted prior to clustering to give greater importance to certain indicators. In the CHSP example, the weights were obtained objectively by modelling the continuous variables of interest. The resulting groups combine estimation domains that are affected by similar quality issues for a given estimate. This provides users with a description of the quality components in addition to the CQI value and a label to report on the estimate's accuracy. The CQI was first released in September 2021 in selected tables in the CHSP. This is the first time that a rating representing the accuracy of estimates has been offered to users as part of a statistical program based entirely on the integration of administrative data at Statistics Canada.

References

Amaya, A., Biemer, P. and Kinyon, D. (2020), "Total Error in a Big Data World: Adapting the TSE Framework to Big Data," Journal of Survey Statistics and Methodology, 8 (1), pp. 89–119. https://doi.org/10.1093/jssam/smz056

Reid, G., Zabala, F. and Holmberg, A. (2017), "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ," Journal of Official Statistics, 33 (2), pp. 477–511. http://dx.doi.org/10.1515/JOS-2017-0023

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

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

Statistics Canada (2019), Statistics Canada Quality Guidelines, Sixth Edition, Catalogue no. 12-539-X, Ottawa, Ontario. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm

Statistics Canada (2021), *Table 46-10-0027-01 Residency participation of residential properties, by property type and period of construction*. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=4610002701&request_locale=en

Statistics Canada (2021), *Table 46-10-0053-01 Ownership type and property use by residential property type and period of construction.* https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=4610005301&request_locale=en

Statistics Canada (2021), *Table 46-10-0054-01 Residency ownership and property use by residential property type and period of construction* https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=4610005401&request_locale=en

Zhang, L.-C. (2012), "Topics of Statistical Theory for Register-Based Statistics and Data Integration," Statistica Neerlandica, 66 (1), pp. 41–63. http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x