SAS® GLOBAL FORUM 2019

USERS PROGRAM

APRIL 28 – MAY 1, 2019 | DALLAS, TX

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

Peter Timusk

Please use the headings above to navigate through the different sections of the poster

Many generations of statisticians have studied survey data and the art and science of conducting surveys. Techniques have been developed that can indicate the quality of a survey estimate. Similarly, work continues on defining quality indicators for administrative data, such as the proportion of missing values of a variable. Big data is a new area with little study on how quality is defined. This poster explores quality indicators in these three data source domains.

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

CENTRE FOR SPECIAL BUSINESS PROJECTS | CENTRE DES PROJETS SPÉCIAUX SUR LES ENTREPRISES

Statistics Canada | Statistique Canada

Please use the headings above to navigate through the different sections of the poster

Intro

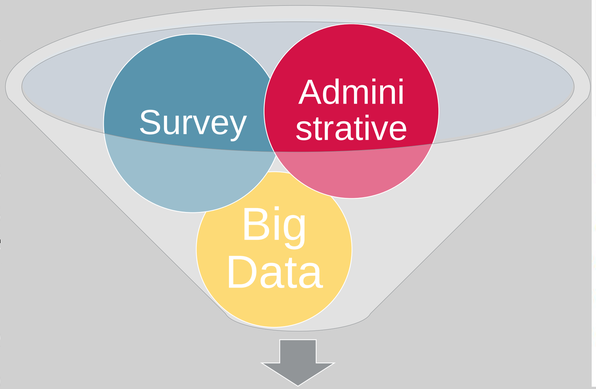
- The use of data to inform decisions is increasing.
- The use of the Internet and transactions online are creating vast amount of data.
- The quality of data is important for its use:
  - **Fitness for use**
  - **Free of bias**
  - **Free of error**
  - **Measures what it is supposed to measure**

Objective

To indicate to data users the quality of the data we are providing.

We have used survey data for decades and have studied it's quality indictors extensively. Example: level of standard error (SE).
We won't address survey data here but it is the standard we would like to arrive at with other data sources.

Sources of data and fitness for use



Fitness for use

Standard error indicators of quality of a statistical estimate

		Canada (map)	
		All surveyed industries	
Enterprise size[3]	Process innovation expenditures[1]	2009	2012
	$150,000 to $499,999	13.7[B]	23.8[E]
	$500,000 and more	37.3[E]	39.2[E]
Large enterprises (250 and more employees)	No expenditures	5.3[B]	2.5[A]
	$1 to $49,999	1.5[A]	11.4[B]
	$50,000 to $149,999	F	25.4[B]
	$150,000 to $499,999	F	11.4[A]
	$500,000 and more	F	49.2[B]

Symbol legend:

A : data quality: excellent
B : data quality: very good
E : use with caution
F : too unreliable to be published.

Source: Survey of Innovation and Business Strategy 2014, Statistics Canada.

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

CENTRE FOR SPECIAL BUSINESS PROJECTS | CENTRE DES PROJETS SPÉCIAUX SUR LES ENTREPRISES

Statistics Canada | Statistique Canada

Please use the headings above to navigate through the different sections of the poster

Administrative Data

- Administrative data can replace survey responses
- Can sometimes be considered a census rather than a sample and therefore improve data quality.
- Data originally collected for other purposes
- **Tax data provide typically**
  - A business's revenue
  - A business's number of employees

Big Data

- Can be gathered **without** discrimination or filters and that may degrade it's fitness for use.
- Is not always a sample but can be sampled.
- Finding a needle in a haystack.
- Quality understudied at this point compared to surveys but many statistical agencies have begun studying this.
- Automated data gathering such as sensors in roads can improve quality.
- Privacy concerns need to be addressed. Legality of data gathering is questioned.

Working definitions of statistical quality and/or accuracy

- All units (people, business, landscape) show up on a file
- A complete and current well classified list of units exists
- The National Statistical Office can have access to all data
- All units can be matched without error
- Target concepts are equal to data concepts
- All information is provided (no nonresponse)
- No measurement errors or bias
- Links between variables are high
- Standard errors are known.

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

CENTRE FOR SPECIAL BUSINESS PROJECTS | CENTRE DES PROJETS SPÉCIAUX SUR LES ENTREPRISES

Statistics Canada | Statistique Canada

Please use the headings above to navigate through the different sections of the poster

Administrative Data

- Taking tax records as an example of administrative data.
- All taxpayers and how to measure how close to ‘all’ the tax records are.
- For personal taxes the social security records and other cross referencing systems suggest a well classified list of units.
- Legal agreement to share tax data with agency.
- Matching is not always possible. Incomplete names, or addresses or SIN numbers. Business registry numbers.

Suggested quality measures, working definitions.

1. All units (people, business, landscape) show up on a file
2. A complete and current well classified list of units exists
3. The National Statistical Office can have access to all data
4. All units can be matched without error
5. Target concepts are equal to data concepts
6. All information is provided (no nonresponse)
7. No measurement errors or bias
8. Links between variables are high
9. Standard errors are known.

Further discussion of measures as they apply to administrative data

- Only provides information that tax records provide and may need to be joined to survey data to get to target concepts.
- Not all fields in tax forms are mandatory so much data is missing (nonresponse).
- How do we determine an indictor for measurement error or bias.
- Accounting rules provide links between variables in tax data.
- Not a survey, so the idea of randomization is not immediately present and need to sample the administrative data. Standard Error not calculated.

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

CENTRE FOR SPECIAL BUSINESS PROJECTS | CENTRE DES PROJETS SPÉCIAUX SUR LES ENTREPRISES

Statistics Canada | Statistique Canada

Please use the headings above to navigate through the different sections of the poster

Big Data

- There is no one example data type discussed here. This could be sales transaction records at a retail chain store or a day of tweets collected from users who report the USA as their location.
- All units may not show up on a file and we may not know how many are missing because we may not know the total number of units.
- Known shoppers at a retail store should be a complete file but the store may not record names and addresses and thus a classified list may not exist. No real validation and classification on Twitter accounts. Could use only verified Twitter accounts. Still work needed.

Suggested quality measures, working definitions.

1. All units (people, business, landscape) show up on a file
2. A complete and current well classified list of units exists
3. The National Statistical Office can have access to all data
4. All units can be matched without error
5. Target concepts are equal to data concepts
6. All information is provided (no nonresponse)
7. No measurement errors or bias
8. Links between variables are high
9. Standard errors are known.

Further discussion of measures as they apply to big data

- These tend to be private sources of data and a government statistics office may have no special reach to obtain the data.
- Transaction records may not be matchable to other records for lack of identifier. Only fields are sales amount, and item, and store location and not customer ID perhaps. Twitter accounts can be fake and not matchable.
- Retail items vary in size, weight, quantity etc.. Subjects of tweets can vary considerably.
- All information is provided perhaps with retail records. Perhaps issues collecting all tweets. Perhaps measurement errors or bias could exist as to types of retail stores or tweeters being a biased population of politically engaged citizens or not.
- Links between variables could be obscure in natural language files like tweets or very clear in transaction records.
- Standard errors may not be valid concept here. Introduction of normal distributions, so classic statistics can apply.

Survey Data, Administrative Data, and Big Data: An Exploration of Quality Indicators

Peter Timusk

Centre for Special Business Projects, Statistics Canada

CENTRE FOR SPECIAL BUSINESS PROJECTS | CENTRE DES PROJETS SPÉCIAUX SUR LES ENTREPRISES

Statistics Canada | Statistique Canada

Please use the headings above to navigate through the different sections of the poster

Conclusion

We have only explored some antidotal thoughts about the various classical statistical data quality metrics that may exist or occur in big data and administrative data which both start outside classical sampling statistics theory. The work developing the theory to bring in administrative data sources is decades old now. In the author's daily SAS programming administrative data such as enterprise tax records are routinely linked to business lists and survey data. The use of big data is under exploration at statistical agencies around the world. Both administrative data and big data will need quality indictors to be useful in the future. Having these indictors available will in itself be sign of quality of these sources.

References

1. Example of Standard Error indicators of quality of a statistical estimate from the Survey of Innovation and Business Strategy, 2014, Statistics Canada.
2. Working definitions of statistical quality and/or accuracy, 2019, Internal discussion document at Statistics Canada.