Providing high quality statistics

High Level Seminar on integrating non-traditional data sources in the National Statistical Systems

Santiago, Chile, October 1-2, 2018

There is no well-established quality framework for statistics based on Big Data

- Statistics based on Big Data sources is still a young field, and the adaptation (or creation of a new) quality framework needs time.

- Big Data sources are so diverse, that it is hard to cover all quality aspects in one framework.

- Because of the large volume of data, big data is generally processed outside the statistical office.

Six criteria for quality in statistics

- Relevance

- Accuracy

- Timeliness and punctuality

- Accessibility and clarity

- Comparability

- Coherence

Relevance

- Do the statistics meet current and potential users' needs?

- Are all the needed statistics produced?

- Do the concepts used (definitions, classifications, etc.) reflect user needs?

- Do all statistics produced have users?

Timeliness and punctuality

Timeliness:

- Is the time lag between the availability of information and the event or phenomenon it describes acceptable to users?

- Do users often quote other sources, rather than the national statistical office?

- **Punctuality:**

- Is there an official data release calendar ?

- Are data normally delivered on the target date?

Accessibility and clarity

- Are key data published regularly and widely?

- How easy is it to find and download or order the data?

- Are the data accompanied by appropriate definitions and explanations (metadata) and information on their quality (including limitations on how the data can be used)?

- Is there a contact point where additional assistance can be provided by the NSI?

- Is data available free of charge, or is there a clear pricing policy?

Accuracy

- Are the methods used to estimate or calculate statistics well established and adequate?

- Are the primary data checked for errors?

- Is the sample size satisfactory?

- If administrative data or non-traditional data sources are used, are they adequate for the purpose?

Comparability

- ***Comparability over time***: Are the data for different periods compiled in the same or similar way so that results can be properly compared over time?

- ***Between geographical areas***: Can the data compiled for different regions be compared with each other?

- ***Between domains***: Are the data for different domains compiled in such a way that results can be properly compared with each other, for example between industrial sectors, between different types of households, different modes of transport, etc.

Coherence

- Can the data be reliably combined in different ways and for various users?
- It is easier to show cases of incoherence than to prove coherence

Experience of the pilot projects

Seven aspects of quality identified:

- coverage

- comparability over time

- processing errors

- process chain control

- linkability

- measurement errors

- model errors and precision

Quality criteria

Traditional

- *Relevance*
- *Comparability*
- *Accuracy*
- *Timeliness and punctuality*
- *Accessibility and clarity*
- *Coherence*

Non traditional

- *coverage*
- *comparability over time*
- *processing errors*
- *process chain control*
- *linkability*
- *measurement errors*
- *model errors and precision*

Findings

- Many causes of error were found
- Data sources may change over time
- Clear need for big data specific checks and correction methods
- Technological changes
- changes in the policy of the data holder
- changes in the population composition and/or amount included.

Conclusion

- Big Data quality has some familiar aspects and some new aspects.

- Diverse nature of Big Data sources makes it difficult to apply standardised quality measures for different projects.

- The current quality framework needs to be extended to better cover Big Data.

For more information

- ESSnet Big Data (2018) Report describing the quality aspects of Big Data for Official Statistics

- UNECE, (2013) What does "big data" mean for official statistics

- UNECE (2014) A Suggested Framework for the Quality of Big Data

Last, but not least

- European Conference on Quality in Official Statistics
- Three day conference, plus one day of training courses
- Every two years
- There is a fee for participation.
- Q2018 was held in Krakow, Poland
- Next Q conference will be in 2020

Thank you for your attention

konstantinos.giannakouris@ec.europa.eu