Session 3 | Quality for non-traditional sources

Discussion

Natalie Shlomo

University of Manchester, United Kingdom

The University of Manchester

Topics

- Overview of papers
- Commonalities and Observations
- Questions for discussion

Paper 1: Quality Aspects Using Mobile Network Operators data for Official Statistics

- Need to ensure standardization and quality assessment of MNO data
- Quality assessment needs to include stages of preparing MNO data for use in the organization, including the pre-processing within the MNO and ingestion of processed data by the NSI
- Quality framework in-line with standard ESS quality framework, including quality at the institutional level and input level, with a focus on hyper-dimensions as developed for the quality framework of administrative data (Daas et al. 2011)
- For official statistics, there is a need to ensure steady streams of data with reliance on private companies: new forms of legislation or firm contract agreements in place
- Focus in the paper on conceptual and qualitative descriptions of quality at different levels but no quantitative metrics proposed
- Given the pre-processing requirements within the MNO, one output quality dimension not mentioned is the need for transparency

Paper 2: Navigating Quality Challenges in Landscaping Web Data: New Aspects and Source Stability

- Paper does not set out a standard quality framework rather focuses on different adaptations of web-scraping depending on project objectives: (1) gathering information through web-scraping for an existing frame of units or (2) first identifying target population and web-scraping of identified websites
  - The latter (2) will likely suffer from selection biases
- Selecting sample of URLs for web-scraping can be done randomly for (1) but for (2) need to consider forms of quota/cut-off sampling
- Authors describe good use of ML methods to facilitate selection of URLs for web-scraping
- Once data is ingested, there is a need to adapt the standardized ESS quality framework at different dimensions (and- hyper-dimensions (see paper 1))

Paper 2: Navigating Quality Challenges in Landscaping Web Data: New Aspects and Source Stability

- Test case based on online job advertisements with websites determined from Eurostat based on an overall score developed from:
  - Analytic Hierarchy Process (AHP) score
  - Popularity, stability, coverage (ICE) rank
  - Website is scraped for small scores

- New use test case based on selecting websites systematically according to checklist of characteristics of the website
  - This requires good meta-information of the website
  - If there is a ‘captcha’ or blocks robots, the website is immediately rejected
  - Must have mandatory set of variables
  - Characteristics 0/1 aggregated and high-scored websites scraped

Paper 3: Assessing the Quality of Transaction Data for use in the Consumer Price Index

- More experiences using transaction data across NSIs, particularly due to impact of pandemic
- Quality assessment of the data in-line with standard ESS quality framework
- Testing of the ingested data and checks over time revealed some problems (eg. duplicated data) and hence need to continue with this check in the production pipeline
- Would have liked to hear more about the test where both manual and transaction data collected at the same time and any quality issues arising
- Authors state that there are ‘no issues in accuracy’ given all transaction data is ingested, but this still needs to be quality assessed given sudden shocks and systematic (informative) missing data
- Risk mitigation if the stream of transaction data suddenly stops and what agreement processes are in place to ensure ongoing data ingestion from private stores
- There were no quantitative quality metrics , eg. more description of the ‘metric analysis program’

Commonalities and Observations

- The main concern is that given a production pipeline application, data ingestion from private organizations can be stopped so risk mitigation needs to be in place
- More work need to be done to understand the representativeness of new forms of ingested data coming from private companies
  - For official statistics purposes, transaction data and web-scraping may be better suited when gathering information on existing units in frames
  - For MNO data, vital to understand the coverage of the operator and to compensate for errors
- The quality assessment needs to include checking the streamed data longitudinally in a systematic manner (not just for testing purposes), i.e. producing quantitative metrics over time comparing current and past injected data

Commonalities and Observations

- Distinguish between one-off checks and ongoing checks for transactional data/streamed data, producing time-series of quality metrics to detect sudden shocks and missing data

- More research needed into quantitative metrics and producing a quality framework scorecard to produce a dashboard for each new form of data and its product, for example:
  - The conceptual and qualitative quality descriptions in papers 1 and 3 can be put into likert-type scales within quality dimensions, and a PCA can be carried out to obtain a quantitative score
  - Ingesting web-scraped data will also need to fit into the ESS quality framework

Questions to Authors

Authors to:

- Comment on the points raised from each paper;
- Address where the research is heading;
- Do they foresee actual applications in the production pipeline of official statistics?

Discussion