Statistics Poland

Information quality frameworks and data stewardship in official statistics

Q2022 Conference

Dominik A. Rozkrut

2022-06-08 - Vilnius

Fundamental Principles of Official Statistics

Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation.

To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Thesis and objectives

- Thesis
  - There’s a need for an integrated comprehensive framework for quality management in official statistics
- Objectives
  - reviewing the current state and identifying areas in need of development work
  - taking into account the observed trends in the information environment of official statistics and its potential role

Information environment of official statistics

Statistics Poland

Information environment

- The information environment of the society and the economy influences public statistics
- Public statistics can positively influence the quality of information environments in contemporary socio-economic systems
- (Olenski 2020)

Information environment

- All resources, processes and information systems, legal norms, and information entities constitute the information environment of people and establishments participating in information processes and systems.
- The information environment determines the scope of information, semiotic forms, and organizational, legal, technical, and economic conditions of access to information by specific groups of people and socio-economic entities.

Challenges

- Exploding Data Ecosystems
- Skyrocketing Number of Actors
- Overwhelming Complications of Complexity
- Lacking Skills, Competencies, Resources
- Insatiable Demand for Insight

Paradigm change

- Scarcity vs. abundance of data points:
  - sampling vs. overrepresentation
- Raw data (e.g., transactions) vs. processed data (e.g., survey):
  - raw is not what it used to be
- Secondary data vs. primary data:
  - data reuse, data stewardship

Data stewardship

- The role of official statistics goes beyond conducting production of statistics and disseminating their results, and
- now extends to constituting information standards
- that help shaping a secure information environment
- of the society and economy

Roles and responsibilities of Data Stewards in the public and private sectors

- Objective: to enable systematic, sustainable, and responsible re-use of data through cross-sector data collaboration in the public interest

- How: data stewards are empowered to create value by facilitating re-use of data; identifying opportunities for collaboration and responding pro-actively to external requests.

- Responsibilities: collaborate, protect, act

- Roles: partnership and community engagement; internal condition and stuff engagement; data audit, ethics, and assessment of value and risk; dissemination and communication of findings; nurturing data collaborative to sustainability

- (Verhulst., 2018)

Essential roles to be played by the NSIs as a Data Stewards

- Spreading the information culture of the society adequate to the current progress of information technologies, processes, and systems

- Promotion of good, consistent, even obligatory information standards covering the minimum quality requirements to be met by all information appearing in the public space and infrastructure systems, information processes and resources

- Promotion of information systems and resources that provide only high-quality information, available as a public good

- Limiting within the public sphere the use of information systems and services that do not guarantee high quality

- Combatting untruths, false information, rumors, incorrect reporting, algorithmic bias, prevention of information provocations used to influence social behavior

Quality

Statistics Poland

What is quality, anyway?

- qual·i·ty /ˈkwɒlәti $ ˈkwɑː-/ noun (plural qualities)
  - 1. (countable, uncountable) how good or bad something is: “Use only high quality ingredients.”
  - 4. (uncountable) a high standard: “Wines of quality”
- ISO 8402-1986
  - “the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs”
- ISO 9000-2005
  - “the degree to which a set of inherent characteristics fulfils requirements"

Back to the core: estimate and error

- In the pure statistical sense,
  - error is the difference between an estimate and the associated true value.
- Desirable properties of the estimator
  - unbiasedness
  - efficiency
  - consistency

Non-sampling errors

- Coverage error
- Measurement error (poor questionnaire design, interviewer bias, respondent error, problems with the survey process)
- Non-response error (total nonresponse error, partial nonresponse error)
- Processing error (coding errors, data capture errors, editing and imputation errors)

European Statistics Code of Practice

- Official statisticians are just statisticians who know how to gracefully recover from their mistakes
- Quality Assurance Framework of the European Statistical System
- Quality of statistics is defined by Eurostat with reference to the following six criteria:
  - relevance; accuracy; timeliness and punctuality; accessibility and clarity; comparability and coherence

Do timeliness matters?

- Is the quality of the particular estimate lowering over time?
  - Or the accessibility, or clarity? No, it doesn’t (ceteris paribus)
  - Or comparability, or coherence? No, it doesn’t (ceteris paribus)
- Extended concept of quality, includes usefulness
  - Does usefulness change over time? Yes, it does.

Selected aspects of quality management in official statistics

Statistics Poland

Information quality

- The right to the truth, to freedom of speech, thought and expression are indispensable for the democratic societies
- In contemporary societies economies, the quality of the information environments in which people, economic entities, state institutions, and international organizations function impacts the political, social, and economic order
- The fundamental law of information:
  - the worse information displaces better information
- (Olenski 2020)

A typology of inaccurate information

- Access to accurate information important, as is combatting untruths, false information, rumours, incorrect reporting, and algorithmic bias
- Even reliable statistical data is used for disinformation by manipulating the statistical data
- Active fabrication vs. no active fabrication
- Intent to harm vs. no intent to harms
- (OCED, 2021)

A typology of inaccurate information

- “Disinformation refers to verifiably false or misleading information that is knowingly and intentionally created and shared for economic gain or to deliberately deceive, manipulate or inflict harm to a person, social group, organisation or country.”

- “Misinformation refers to false or misleading information that is shared unknowingly and is not intended to deliberately deceive, manipulate or inflict harm to a person, social group, organisation or country. The spreader does not create or fabricate the initial misinformation content.”

- (OCED, 2021)

A typology of inaccurate information

- “Contextual deception refers to the use of true but not necessarily related information to frame an issue or individual (e.g. a headline that does not match the corresponding article), or the misrepresentation of facts to support one’s narrative (e.g. to deliberately delete information that is essential context to understanding the original meaning). While the facts used are true (unlike disinformation) and unfabricated (unlike misinformation), the way in which they are used is disingenuous and with the intent to manipulate people or cause harm.”

- “Propaganda refers to the activity or content adopted and propagated by governments, private firms, non-profits, and individuals to manage collective attitudes, values, narratives, and opinions”

- (OCED, 2021)

The criteria for the quality of information

- A hierarchy of criteria for the quality of information and meta-information may be assumed as follows:
  - truth, compliance with facts;
  - social usefulness of information, i.e., compliance of information norms with civilization norms and with the potential use of data;
  - situational usefulness of information, i.e., the relevance and reliability of the information in specific use situations of particular participants in information processes.
- (Olenski 2020)

Revisiting data stewardship

- In times of digitalization and eruption of data environments and ecosystems, information security should be safeguarded by appropriate development of information infrastructure, comprising both resources and systems created and maintained by the state institutions.

- Among those institutions, official statistics should actively take and play a prominent role by extending its classical spectrum of activities to include those prevailing in the role of the data steward.

Anything about the survey design?

- In official statistics, comparative descriptive statements are made with increasing frequency, if not increasing care

- Statisticians make their observations in specific socio-economic systems. These observations are coded directly into a language with direct rules of interpretation or placed indirectly into a language on the basis of inferential rules.

- The argument against the use of general inferential rules in comparative measurement is that the only appropriate framework for assessing characteristics of phenomena must be derived from the systems in which observations are made.

- -“For specific observations, a belch is a belch (...). But within an inferential framework a belch is an “insult” or a “compliment” (...)”

- (Suppes & Zinnes, 1963), (Osgood, 1965), (Przeworski & Teune, 1970)

The logic of measurement in comparative research (especially within EU, ESS)

- A language of direct measurement requires only a grammar and rules of empirical interpretation invariant across cultures or societies. A language of inferential measurement requires, additionally, general statements defining the meaning of a specific observation in terms of its systemic context.

- A language of scientific measurement, can incorporate the context of specific systems, and if so, how the social context can be introduced into measurement statements without destroying the uniformity and generality of the language of measurement.

- The metalanguage for determining comparability is the language of measurement.

- (Suppes & Zinnes, 1963), (Osgood, 1965), (Przeworski & Teune, 1970)

Establishing Equivalence

- Looking for a measure that is reliable across systems and valid within systems
- Relationships among indicators within systems are the basis of validating the indicators across systems
- An indirect yet effective way of testing equivalence is through theoretical assumptions about the behaviour of indicators in particular systems
- Under theoretical assumptions of varying strength, even instruments composed exclusively of indicators specific to particular system can be shown to be equivalent
- Observations acquire meaning only within a theory; theory logically comes before research activity
- (Suppes & Zinnes, 1963), (Osgood, 1965), (Przeworski & Teune, 1970)

The case of Community Innovation Survey & Survey of ICT

- Space: strong systematic structural variation across countries
- Time: changes in definitions (e.g., manuals) and systems (e.g., technologies)
- The importance of cognitive testing to inform underlying theory, as a cornerstone for measurement

Administrative sources, big data - data quality

- “Challenge 1: statistics teaching should cover data quality issues.
- Challenge 2: develop detectors for particular quality issues.
- Challenge 3: construct quality metrics and quality scorecards for data sets.
- Challenge 4: audit data sources for quality.
- Challenge 5: be aware of time series discontinuities arising from changing definitions.
- Challenge 6: evaluate the impact of data quality on statistical conclusions.”
- (Hand, 2018)

Administrative sources, big data - data quality

- “Challenge 7: explore potential sources of nonrepresentativeness in the data.
- Challenge 8: develop and adopt tools for adjusting conclusions in the light of the data selection processes.
- Challenge 9: explore how suitable the administrative data are for answering the questions. Identify their limitations, and be wary of changes of definitions and data capture methods over time.
- Challenge 10: report changes and time series with appropriate measures of uncertainty, so that both the statistical and substantive significance of changes can be evaluated. The measures of uncertainty should include all sources of uncertainty which can be identified.”
- (Hand, 2018)

Administrative sources, big data - data quality

- “Challenge 11: be aware that administrative data are observational data, and exercise due caution about claiming causal links.

- Challenge 12: be aware of the risks associated with linked data sets and the potential impact on the accuracy and validity of any conclusions. Recognise that quality issues of individual databases may propagate and amplify in linked data. Develop better measures of overall combined data quality.

- Challenge 13: continue to develop statistically principled and sound methods for record linkage and evidence assimilation, especially from nonstructured data and data of different modes.

- Challenge 14: Develop improved methods for data triangulation, combining different sources and types of data to yield improved estimates.”

- (Hand, 2018)

Framework for assessing the Quality of Big Data

- Extensions to existing statistical data quality frameworks needed
- Prepared by the Task Team on Big Data Quality, within HLG-MOS
- A structured view of quality at three macro-phases of the business process:
  - input, throughput, output
- A hierarchical structure composed of three hyperdimensions with quality dimensions nested
  - the source, the metadata and the data
- (UNECE, 2014)

Framework for assessing the Quality of Big Data

- “Three general principles are proposed when evaluating Big Data quality:
  - Fitness for use (is the data source appropriate for the purpose)
  - Generic and flexible (a quality framework such as the one proposed here should be broadvand applicable over a wide variety of situations)
  - Effort versus gain (is the effort involved in obtaining and analysing the data source worth thebenefits gained from the data source)”
- (UNECE, 2014)

Framework for assessing the Quality of Big Data

- Input phase
  - engage in a detailed quality evaluation of a source both before and after acquiring
  - in addition to dimensions used to assess administrative data, additional dimensions were suggested
    - privacy and confidentiality
    - complexity
    - completeness of metadata
    - linkability
- (UNECE, 2014)

Framework for assessing the Quality of Big Data

- Throughput phase
  - four principles of processing are proposed:
    - system Independence
    - application of quality dimensions
    - steady states
    - quality gates
- (UNECE, 2014)

Framework for assessing the Quality of Big Data

- Output phase
  - the Australian Bureau of Statistics Data Quality Framework applicable to reporting, dissemination, and transparency
  - additionally new output dimensions were recommended
    - secondary sources and confidentiality for the hyperdimension source; complexity for the hyperdimension metadata
    - selectivity and predictive power for the hyperdimension data
- (UNECE, 2014)

Framework for assessing the Quality of Big Data

- The dimension of predictive power is a necessary addition
  - sampling theory may not be an appropriate metric for evaluating the utility of metrics
  - predictive power considers the ability to predict a variable of interest, where ‘prediction’ here is being used in the statistical sense of providing some kind of empirical estimation
- (UNECE, 2014)

Quality Indicators for the GSBPM

- Quality indicators that have been developed for the production of statistics from both survey and administrative data sources (ADS),
    - with reference to the different stages of the Generic Statistical Process Model (GSBPM) Version 5.0
    - to understand and manage the quality of the statistical products.
- Quality indicators are mapped to each sub-process of the GSBPM
- (UNECE, 2017)

Beyond CoV: AI and ML -> Clustering

- Used for exploratory data analysis, starting with a first understanding of the structure of data
- Little theoretical understanding of clustering, what is a "good" clustering?
- A wide variety of different clustering methods, with different measures of quality, no clear ground truth to evaluate (excepts when with labeled learning data set)
- A clustering may have different value for different uses

Clustering

- Some solutions include
  - objective utility functions: sum of in-cluster distances, other distances, spectral clustering
  - a restricted set of distributions, like mixtures of gaussians

Clustering Axioms

- Postulate axioms that, ideally, every clustering approach should satisfy
- Defining clustering-quality measure: a function satisfying some properties that make this function a meaningful clustering-quality measure
- Clustering-quality measures axioms: scale invariance; richness; consistency; isomorphism Invariance
- Theorem: consistency, scale invariance, richness, and isomorphism invariance for clustering quality measures form a consistent set of requirements
- (Ackerman, Ben-David, 2008)

Machine learning

- “Under ever-increasing complexity of machine learning models, interpretability is suffering
- The necessity for plausibility and verifiability of predictions made by these black boxes is indispensable
- The research community has recognised this interpretability problem and focussed on developing a growing number of so-called explanation methods
- It is, however, of- ten unclear, which explanation method offers a higher explanation quality”
- (Honegger, 2018)

Machine learning

- An axiomatic framework, Explanation Consistency Framework, which allows comparing the quality of different explanation methods amongst each other is proposed
- It consists of three proxies/axioms for explanation quality:
  - 1. Identity: Identical objects must have identical explanation
  - 2. Separability: Non-identical objects can not have identical explanations
  - 3. Stability: Similar objects must have similar explanations

Machine learning

- Interpretability comprises the three goals of
  - -> accuracy, understandability and efficiency
- The axioms influence each of these goal-dimensions of interpretability and that only by at least partially fulfilling these axioms, explanations of an Explenation Method can be interpretable to users

- However, the axioms are necessary but not sufficient to achieve interpretability with an EM

Knowledge quality, ontologies quality

- Ontologies play an important role for extracting the information and knowledge about the specific domain and present the problem in more understandable form.

- Abundance of literature on quality frameworks for evaluation of knowledge quality, two examples:

  - R. A. Khan, U. Qamar, A. Aslam, P. Saqib and A. Ahmad: Quality Framework for Ontologies Evaluation Based on Structural Characteristics, 2019

  - Silvio Mc Gurk, Charlie Abela, and Jeremy Debattista: Towards Ontology Quality Assessment

Communicating quality
What is the role of communication?

- What is the role of communication in the process of statistical research, what components?
- How does it affect the quality of both the process and product?
- Which aspects/attributes of quality are particularly exposed to (or dependent on) the communication
- Chain of communication in statistical process: from data to decision (policy) making – alternative paradigms: evidence based vs. policy driven

Scientific (statistical) communication

- contents/information (‘message’)
- communicator/transmiter (statistician)
- reciver/audience
- communication channel
- code (transmitter’s/receiver’s code)
- context
- feedback
- noise
- (Maggino and Trapani, 2010)

Quality of communication

- Ensuring high-quality (error-free) transmition of information: inter-/ across phases/stages of the statistical process (data production and analysis); among the stakeholders in the statistical process

- Communication of quality (‘reflected in the TQM principle of ‘participation by all’): communication with user, and other stakeholders

- The logic of methods and the logic of actions: two-dimensional communication-route

- Communication towards ensuring methodological quality of research

- Strategies of actions: translation, brokering, interaction model

- (Okrasa 2018), (Prewitt et al., 2012)

Quality certification

- “The quality procedures for internal and external reports, recommendations and briefs;
- The quality assurance of statistical development projects in which methodologists and business analysts participate;
- The quality assurance of methodological courses taught to statisticians;
- The internal management of the department.”
- (Zeelenberg, Ypma, Struijs, 2018)

Conclusion

- There’s a need for an integrated comprehensive framework for quality management in official statistics

Statistics Poland

Thank you

for your attention!

stat.gov.pl
twitter.com/StatPoland