

# Assessing data quality

By Michael Lyon

Tel: 020 7601 5466

Email: [michael.lyon@bankofengland.co.uk](mailto:michael.lyon@bankofengland.co.uk)

*There is increasing recognition by statistics producers of the importance of measurement and disclosure of information relating to data quality. Documentation of data quality is encouraged in the Bank of England's Statistical Code of Practice. Reflecting this, the Bank's statistics division regularly discusses technical aspects of its statistical outputs in articles in this publication.*

*The Bank is now planning to adopt a Data Quality Framework, which will provide general information on data quality, as well as define specific quantitative indicators to measure quality. The Framework will be based on the dimensions of quality as specified under Eurostat's data quality definition. Data quality information will be provided in the explanatory notes to data outputs and through regular articles in this publication.*

## Background

Statistical agencies, including the Monetary and Financial Statistics Division of the Bank, increasingly recognise the importance placed by users of statistics on the availability of information about the quality of their statistical outputs. Transparency of information on data quality enables users to gain a more realistic understanding of the data they depend on, and can help producers to monitor and become more accountable for the quality of their data.

## Systems of data quality

A number of systems for measurement and reporting of data quality have been introduced both in the UK and internationally. In Europe, the European Statistical System (ESS), comprising the Statistical Office of the European Commission (Eurostat) and member state national statistics institutes, first adopted a statistical quality definition in 2003.<sup>1</sup> This definition involves six dimensions of quality for statistical outputs: relevance; accuracy; timeliness and punctuality; accessibility and clarity; comparability; and coherence. (Table 1).

**Table 1: European Statistical System (ESS) definitions of quality dimensions**

ESS Quality dimension	Description
Relevance	Relevance is the degree to which statistics meet current and potential users' needs. It refers to whether all statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflect user needs.
Accuracy	Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values.

Timeliness and punctuality	Timeliness reflects the length of time between availability and the event or phenomenon described.  Punctuality refers to the time lag between the release date of data and the target date when it should have been delivered.
Accessibility and clarity	Accessibility refers to the physical conditions in which users can obtain data.  Clarity refers to the data's information environment including appropriate metadata.
Comparability	Comparability aims at measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical areas, non-geographical domains, or over time.
Coherence	Coherence of statistics is their adequacy to be reliably combined in different ways and for various uses.

Source: European Data Quality Definition.

Within the UK, the statistical codes of practice maintained by the Office for National Statistics (ONS) and by the Bank each include sections on data quality. The ONS has developed a comprehensive approach to quality measurement and reporting, which includes quality measurement guidelines and definitions of some 11 'key quality measures'.<sup>2</sup>

The Bank regularly publishes articles in this publication discussing specific issues of data quality relating to its own statistics. It also provides additional information (or 'metadata') on the relevance, definitions and sources of published data through the explanatory notes to statistical releases and individual data series on the Statistical Interactive Database.<sup>3</sup> The system of explanatory notes has recently been reorganised under a common framework with consistent headings.

<sup>2</sup> 'Guidelines for Measuring Statistical Quality, version 3.0', by ONS, April 2006.

<http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=13578>

<sup>3</sup> Statistical Interactive Database, Bank of England.

<http://www.bankofengland.co.uk/mfsd/iadb/notesiadb/content.htm>

<sup>1</sup> European data quality definition, Eurostat, 2003.

[http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP\\_DS\\_QUALITY/TAB47141301/DEFINITION\\_2.PDF](http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47141301/DEFINITION_2.PDF)

In 2006, the Bank developed a cost / benefits analysis (CBA) approach to its data compilation, to balance the needs of data users against the compliance costs placed on reporting institutions.<sup>4</sup> Any methodical application of a CBA framework should include information on data quality, to assure users that data outputs remain fit for purpose.

## Data Quality Framework

The Bank now plans to adopt a Data Quality Framework which will define appropriate quantitative data quality measures and summarise data quality guidelines in one document. The Framework will be based on the six dimensions of the ESS data quality definitions.

### *Monetary and financial data collected by central banks*

In defining suitable guidelines and measures of data quality, the Framework will recognise the importance of monetary and financial data collection methods typical of central banks. These differ in some important respects from those used by national statistical agencies, such as the ONS in the UK, in respect of national accounts economic data.

The key difference relates to the sampling frameworks employed by central banks and national statistical agencies in relation to their target populations. For example, the scope of the Bank's authority to collect statistics is restricted principally to banks and building societies, which in the UK form the monetary and financial institutions sector. In common with widespread practice among central banks, the Bank has a very high coverage among this population: all banks and building societies above a certain threshold are required to report the basic balance sheet returns on a monthly basis, and all reporters on a quarterly basis. This system of defining reporting populations according to minimum business-size thresholds is known as 'cut off the tail' sampling, and can be contrasted to random and stratified random sampling approaches.

There is a culture of good compliance with the Bank's statistical reporting standards. Data quality of banking data is good: there is in essence a 100% response rate, and reporters engage positively with the Bank's data cleansing processes.

The two approaches give rise to different types of data errors. In random sampling, sampling error arises because variation in the population will translate into variable sample-based estimates; this is in addition to all other sources of error, termed non-sampling error. Sampling error may be minimised but not avoided through the application of good sample design, and its magnitude can be estimated through statistical analysis.

In the case of 'cut off the tail' sampling, sampling error will be less significant although in the absence of information on non-reporters it is difficult formally to assess. However, other forms of data inaccuracy (non-sampling error) may potentially arise in this sampling framework.

## Publication

The Data Quality Framework is planned for release in 2008, and the intention will be to review it periodically.

The Framework will define certain quantified measurement criteria for data accuracy (through coverage and revisions measures), and will formalise existing practices, for example with respect to reporting the results of the annual review of seasonal adjustment. It will serve as a reference to existing statistical data quality guidelines, and will help to explain some of the trade-offs that can arise between competing dimensions of quality (for example, between timeliness and accuracy of data).

It will also, for the first time, set out criteria that will be adopted internally in the review of data imputation methods. Imputation is used by the Bank in cases of non-reported data, for example because of reduced reporting panel sizes. The objective is to maintain good estimates of statistical outputs, even with a reduced data collection burden on reporters.

Data quality information as defined in the Framework will be made available through the explanatory notes facility or through regular articles in this publication, as appropriate.

<sup>4</sup> 'Cost Benefit Analysis of Monetary and Financial Statistics', Bank of England, 2006.  
<http://www.bankofengland.co.uk/statistics/about/cba.pdf>