Quality and Administrative Data

5.1 Introduction

As mentioned in Chapter 4, concerns about the quality of administrative data are often one of the main barriers to their increased use for statistical purposes. These concerns may or may not be justified, and are often based only on specific aspects of quality, such as timeliness. To properly address these concerns an objective quality management framework is needed; one that considers all relevant aspects of quality, and allows an informed decision to be made.

Many statistical organisations have already put in place some sort of quality framework for data collected via traditional survey methods, but relatively few have extended this approach to cover data from administrative sources[1].

5.2 Defining Quality

The starting point for such a framework is the definition of quality itself. Again, much work has been done in this area by national and international statistical organisations, most of which is based on the international standard ISO 9000/2005[2], which defines quality as:

"the degree to which a set of inherent characteristics fulfills requirements."

Unfortunately this definition is not particularly easy to understand, and needs some further explanation. It can be split into the following parts to aid interpretation:

1) "requirements"

This is usually taken to mean the requirements of the user of specific goods or services, though it could also be argued that the requirements of the producer, or even of society as a whole, should also be taken into account. For example a fast car with a large engine may fully meet the requirements of an individual, but may not meet the requirements of society regarding pollution or road safety. However, the ultimate products of official statistical agencies, the statistics themselves, are usually produced within the public sector as a "public good", so in this case the requirements of these different groups are largely overlapping.

If we consider an administrative data set as a product in its own right though, there can be considerable divergence between the requirements of the producer (e.g. an administrative agency) and the user (a statistical organisation). Furthermore, as the "transaction" is often not on market terms, there may be little incentive for the producer to consider the user requirements. This can result in tensions over quality, which underlines the requirement for a sound organizational framework as described in Chapter 3.

2) "a set of inherent characteristics"

Users of any goods or services judge quality against a set of criteria concerning different characteristics of those goods or services. This is often done sub-consciously, as in the example of a meal in a restaurant: An individual will judge the quality of that meal in terms of the way the food was cooked and presented, the quantity, the service, the decoration and ambiance of the restaurant, and perhaps several other criteria (for the moment cost is not included, we will return to this later in the chapter). The quality of statistics can similarly be judged against a set of inherent characteristics or criteria.

Several statistical agencies have developed lists of criteria for evaluating quality of statistical data, however the main international agencies have now reached agreement on the following list:

- Relevance - the degree to which statistics meet the needs of current and potential users. Relevance therefore refers to whether the statistics that are needed are produced, and whether the statistics that are produced are needed. It also covers the extent to which the concepts used (definitions, classifications etc.) reflect user needs.
- Accuracy - the closeness of statistical estimates to true values.
- Timeliness - this reflects the length of time between data being made available and the event or phenomenon they describe.
- Punctuality - the time lag between the date that data were actually released and the target (often pre-announced) release date.
- Accessibility - the physical conditions in which users can obtain data: where to go, how to order, delivery time, clear pricing policy, convenient marketing conditions (copyright, etc.), availability of micro or macro data, various formats (paper, files, CD-ROM, Internet...), etc.
- Clarity / interpretability - whether data are accompanied by sufficient and appropriate metadata, whether illustrations such as graphs and maps add value to the presentation of the data, and whether information on data quality is available.
- Coherence / consistency - data from different sources, and in particular from statistical surveys of a different nature and/or frequency, may not be completely coherent in that they may be based on different approaches, classifications and methodologies. They may not, therefore, convey a completely coherent message to users, e.g. users may be confused if two different measures of the same variable are published with different values.
- Comparability - the extent to which differences between statistics are attributed to differences between the true values of the statistical characteristic, or to methodological differences. Comparability includes:
  - Comparability over time – the extent to which data from different points in time can be compared.
  - Comparability through space – the extent to which data from different countries and/or regions can be compared.
- Relevance - the degree to which statistics meet the needs of current and potential users. Relevance therefore refers to whether the statistics that are needed are produced, and whether the statistics that are produced are needed. It also covers the extent to which the concepts used (definitions, classifications etc.) reflect user needs.

This list of criteria can be used in two ways relating to administrative data. Firstly it can be used to assess the quality of the resulting statistics, and to compare data based on administrative sources with those based on surveys. Secondly, the list can be used to help evaluate the quality of different administrative sources themselves[3]. For example, if a statistician is fortunate enough to be faced with a choice of two or more administrative sources, it can help to determine which source has the higher quality.

However, if the list is being used to assess the quality of administrative data, it should be noted that absolute accuracy can be difficult to determine if there is not sufficient supporting information about the population and the collection process. In this case, two factors should be considered, the credibility of the source and the plausibility of the data, i.e. whether the source is trusted, and whether the data look reasonable when compared to other sources, and to the values the statistician would expect. For a more objective measure, some sort of quality survey may be needed to determine the correct values of certain variables.
The closeness of administrative units and variable to the units and variables required for statistical purposes can be an important factor in determining the quality of an administrative source. The fewer transformations required, the lower the risk of error or bias. This aspect can be considered as part of the criterion of coherence.

5.3 The Constraint of Cost

Cost is deliberately excluded from most lists of statistical quality criteria, as it is considered to be more of a constraint. Once quality has been determined, cost is added to the equation to allow practical decisions on cost-efficiency to be made.

Cost is, however, particularly important in the case of administrative sources, because where they are shown to deliver a lower absolute level of quality than survey data, they may still have a sufficient cost advantage, which could make them the most cost-efficient option. It may also be possible to channel some of the cost savings into improving quality, thus reducing or eliminating the quality gap.

5.4 Quality Measurement in Practice

To fully understand the quality of administrative sources, and their impact on the quality of statistics, we need to consider three elements:

1) The quality of incoming data

The incoming data, whether they are from administrative or survey sources, can be judged against set of criteria such as those listed above. The most important criteria are likely to be timeliness, and relevance in terms of the extent to which the coverage and concepts of the source meet requirements. Comparability with other sources can also be important, and some sort of exercise to reconcile data from different sources may be necessary from time to time to get a clear picture of quality. Quality check surveys are sometimes used for this purpose.

One point worth bearing in mind is the extent to which the data subject has an interest in the quality of the data. The amount of effort and care put into providing the data will vary according to the perceived value or importance of the data collection, thus data subjects may, in some cases, provide better quality data for administrative purposes than they do for statistical purposes.

2) The quality of data processing

Even if the incoming data are perfect, their quality can still be affected by the different processes they go through before they are used for statistical outputs. Ideally processing should improve quality, but unfortunately this is not always the case. Examples of how data processing can affect quality include:

- Data matching and linking – too many false matches will lead to errors in the data; too many false non-matches will lead to duplication, which will overstate the size of the population of interest, and possibly introduce bias.
- Outlier detection and treatment – using outlier detection methods to detect errors can help to improve the quality of the data, and generally the more extreme the outlier, the more likely it is to be an error. However over-zealous treatment of outliers will result in genuine data values being altered and can lead to important trends in the data being missed.
- Quality of data editing – as for outlier detection and treatment, data editing should improve quality, but if not done carefully it can introduce error and bias.[4].
- Quality of imputation – if imputation is used to fill missing values or records it can help to improve coverage, but again the methods used need careful scrutiny to avoid the introduction of bias.

One very important principle that should always be followed, particularly when processing data from administrative sources, is to keep a copy of the raw data (and any associated metadata) to refer back to if necessary. Comparisons of data before and after processing can help to assess the quality of that processing, and to identify any specific problems.

3) The quality of statistical outputs

The usual interpretation of the ISO quality definition by statistical agencies is that quality is all about meeting user requirements. The quality of statistical outputs is therefore determined in this context. This means that it is necessary to determine these requirements, to discuss them with users, and to get regular feedback, for example via user satisfaction surveys.

Moving from survey to administrative sources will clearly have an impact on output quality. Typically this impact may be positive for some quality criteria, and negative for others. In all cases, it is necessary to get an overall view of the impact, giving greater weight to those criteria the users consider to be the most important. For example, users may feel that an improvement in timeliness more than compensates for a reduction in accuracy, particularly for short-period economic data. Another consideration should be the impact on time-series data, and whether it is possible to construct a consistent series of sufficient length following the change.

It can be particularly important to give at least as much weight to the views of users as to the perceptions of statisticians, which may, in some cases be too heavily focussed on traditional notions of accuracy. Overall, it is vital that any judgement of the impact on statistical outputs is based on objective evidence rather than on supposition, as this is the only way to counter the potential for resistance to change as described in Chapter 4. One way to ensure this is to use quality reports, following standard templates,[5], to document and communicate the impact of changing data sources.

5.5 The Role of Metadata

Metadata[6] are vital for informing both producers and users about data quality. They should be present at all three of the stages referred to in the previous section. Incoming data should be accompanied by sufficient metadata to fully understand them, and to ensure that values are correctly allocated to the relevant variables. Detailed documentation on the concepts, definitions and purpose of the source, as well as on the collection and processing methods used, is also important. This will give a better understanding of potential quality issues, and should form the basis for data editing rules in the processing stage.
During data processing it is important to record what has been done to which records and values. This not only provides vital information for assessments of processing quality, but also provides a mechanism to investigate any potential problems in the process and undo any errors.

Statistical outputs should be accompanied by sufficient metadata to allow users to retrieve them, interpret them correctly, and form an opinion on their quality. For regular and heavy users of the outputs, full documentation of all three stages, preferably following a standard format, will provide the necessary information to enable them to draw the correct conclusions from the data. Communication of quality can often be difficult to get right, as some users want full details, whereas others are happier with very high-level summary indicators. A metadata model that allows users to see different levels of information, starting with a summary, but with an option to see greater detail, is perhaps the most appropriate.

5.6 Summary

The best way to assess the quality of an administrative source is to build up a thorough knowledge of that source, including the primary purpose of the source and the way the data are collected and processed. Thorough understanding of a source will allow a more accurate assessment of strengths and weaknesses.

To assess the impact of using different sources, it is necessary to combine knowledge of the sources and the processes used to convert them to statistical outputs with the views of the users of those outputs. This will then allow an objective and holistic assessment of the impact of using administrative data versus statistical survey data.

[3] For an application and extension of this approach, see the discussion paper from Statistics Netherlands "Checklist for the Quality Evaluation of Administrative Data Sources": http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf
[4] For a comprehensive collection of papers on different data editing issues, see the working papers of the Statistical Data Editing Work Sessions organised by the UNECE - http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing