Scottish Social Services Council

A Guide to Basic Quality Assurance in Statistics

Quality assuring data

The production of high quality statistics is a fundamental priority for the Scottish Social Services Council (SSSC). There are frequent and important statements about quality in the Code of Practice for Statistics and these are picked up in the National Statistician's guidance on Quality on the UK Statistics Authority's website. These are clear about the requirement on those producing Official statistics to ensure that quality is monitored and assured.

The definition of Quality as relevant to Official Statistics is necessarily broad and covers:

- relevance
- accuracy
- timeliness and punctuality
- accessibility and clarity
- comparability
- coherence.

This guide is not intended to be exhaustive and should not be treated as a check-list; rather it should help you to raise questions about the approach you are taking to the quality assurance of your data. Reflecting on your current approach may help you to put in place necessary improvements and the implementation of a robust approach to quality assurance which remains valid over the coming years.

Contents

1. Planning of data quality assurance

1.1 Quality assurance (QA) of the data we use is a fundamental part of ensuring that:

- data is used in an appropriate way
- the risk of errors in our statistics is minimised.

1.2 Errors in published statistics can have serious consequences:

- important decisions may be taken based on incorrect information
- the trust of internal and external customers in statistics produced may be affected
- an error may have a significant political impact (for example, if a correction affects the achievement of a government target).

1.3 If an error does occur, correction of the error is likely to require a considerable amount of extra effort from the statistical staff involved, and others. Correction of errors should adhere to the SSSC corporate revisions policy and steps to take may include:

- correcting the publication online
- adding an explanation of the error to the website
- informing Ministers and other customers of the mistake.

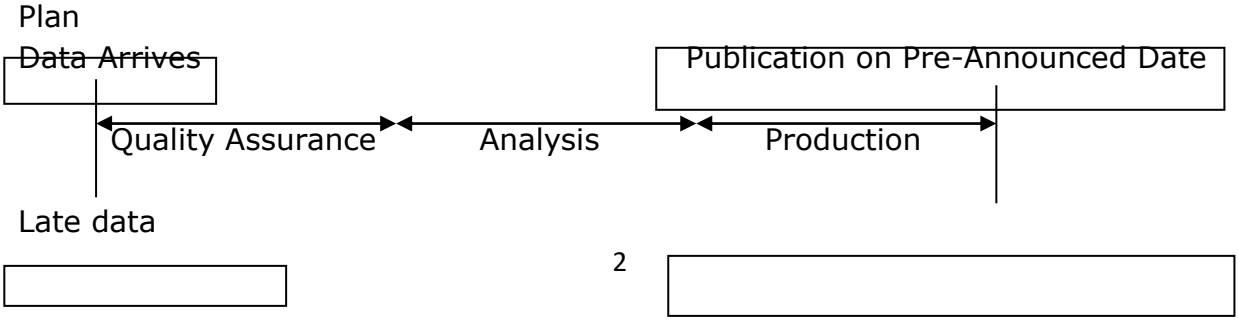
1.4 Although there is always some risk that errors may occur, it is very important that sufficient time is set aside to quality assure the data before analysis work is started. There should be a clear plan of the quality assurance that is to be carried out, often drawing on previous experiences, and this should be used to decide on the time required (ideally putting in some slack in case of delays).

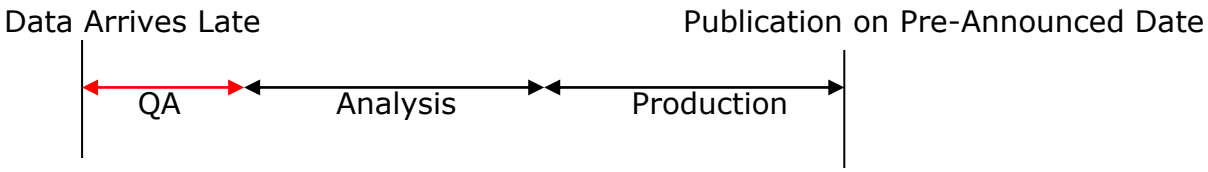
1.5 A review of the processes and problems encountered last time the data were analysed and published may help prioritise work and may prevent the repetition of mistakes.

1.6 It is worth discussing what issues there were last time a piece of work was carried out, and what the impact of these issues was. Input from analytical, policy and other colleagues at the start is better than being told at the end 'we always have problems with that data, with the interpretation, with the policy impact etc.'

1.7 It is also important to make it clear to others involved why the QA process is important and how long it is expected to take.

1.8 Official and National Statistics have pre-announced publication dates. However, it is important to avoid the situation where, if the data arrives late and, in order to publish the statistics on time, quality assurance checks are reduced (as illustrated below), increasing the risk of errors.





1.9 To minimise the risk of data arriving late, it is important not to just expect the data to arrive on the date originally agreed with the data provider.

1.10 It may be helpful to keep in touch with the data provider on a regular basis to ensure that there are no unforeseen problems that may affect when the data will be received (for example, due to technical problems or lack of staff).

1.11 It may also be possible to see an early version of the data to obtain a feel for the format the final dataset will arrive in, and possibly carry out some basic checks.

1.12 For external contractors it is also worth checking to see if there is a contractual agreement specifying when the data will arrive and what format it will be in.

1.13 If data are received late, options include:
pulling in extra people to enable the checks to be carried out in a shorter period (where resources are available) delaying the publication (allowed if there is a very clear, non-political, reason – for example, if it is unavoidable that data will arrive late from external data providers). The Office of the Chief Statistician can advise on the steps to take if a publication date is to change.

2. Basic checks

2.1 These are some basic ways of quality assuring data prior to analysis which can help to identify any problems in the data. These are some of the steps that can be used to check the data.

2.2 Preliminary checks
2.2.1 It is important to understand what data kind of data is expected. Is it raw unit level data, aggregated data, processed data etc. – this will aid understanding of the kinds of quality assurance required.

2.2.2 Check that the dataset is complete. Are there any missing values? Check against expected number of data entries where possible – e.g. total number of schools, Scottish population, survey respondents, etc.

2.2.3 For survey data – check number of responses is correct based on routing (e.g. follow up questions asked only of those who answered yes to a previous question). It is also worth checking all variables are present.

2.2.4 Also check the dataset for possible duplicate entries, which may be the result of a mistake in the data entry.

2.3 Data checks
2.3.1 Scan through the dataset and consider if the values in the dataset look sensible (e.g. if the Census reported the population of Scotland to be 10 million, it would be apparent that the value is not sensible).

2.3.2 Plotting the data on a chart is a helpful way to identify outliers in the data and helps to identify the trend of the data. Compare the new data with previous years, has there been any large changes between years. If so, do you know a reason why? For example a change in the methodology or change in policy intervention. It is important to ensure possible reasons are clearly evidenced and not just assumptions.

2.3.3 Check that the values in columns and rows add up to the totals. Cumulative totals in columns and rows are very useful to check the data is accurate and does not contain mistakes (e.g. the area of derelict land in each local authority figures, when added together, should equal the area of derelict land in Scotland figure).

2.3.4 If possible, compare data with previously published data to check they correspond as some data sources may have already published similar data. Consider similar publications from the past but also other products (compendium publications, PQs, similar analyses produced by other organisation etc.) so you can be confident that your data are sensible. This is particularly important when preparing compendium publications. As with all quality assurance if the data do not correspond, consider why not and then take appropriate action.

2.3.5 Ensure that all figures reported throughout a publication are consistent (i.e. if reporting a value in Table 2 of a publication then again in Table 4, make sure these numbers correspond). Also ensure that numerators and denominators are correctly applied throughout and between tables. Consider the conclusions drawn from the basic data and those from more processed data such as proportions, summary statistics and charts – are they the same, if not, why not?

2.3.6 Consider the 'common sense' of the final figures. Perhaps consider links to other data, for example expenditure data. This might be particularly relevant when considering changes or differences between groups (sectors in the economy, types of farming, local authority data, gender analysis etc.).

3. Is the dataset fit for purpose?

3.1 It is important to become familiar with your data and use it in the correct manner. There will always be occasions where statistics are either misused or

quoted out of turn, but in striving to "meet customer needs by producing relevant and reliable information, analysis and advice, free from any political interference" the chances of this happening are considerably reduced.

3.2 In trying to carry out a specific type of analysis it is important to ensure that the data collected is suitable for the piece of analysis for which it is intended. The following points should help ensure that data we publish is not vulnerable to being misinterpreted.

3.2.1 Consult with colleagues, stakeholders, users etc. to gauge the need and practicality of the piece of analysis planned. It could be the case that, without realising it, their needs can actually be met with data already collected.

3.2.2 Scrutinise the data currently collected in order to be aware what it can and cannot be used for. Key issues when studying data include:

- what is the nature of the data, i.e. is it performance measuring, context/outcome based etc
- the source of the data – how has it been collated
- is the dataset complete. If a complete dataset cannot be obtained what is the best method for estimating/forecasting missing data. If the data is weighted, perhaps focus should be placed on data quality for the components with the largest weights
- what timescale does the data cover and has the collection process been consistent throughout this period
- will any outliers effect/skew the analysis being carried out, and if so is it appropriate to remove outliers etc
- what does the data plot show. Would a transformation of the data be important to enable you to carry out the analysis. Are there any seasonal patterns that will require seasonal adjustment/smoothing to be carried out
- is the sample size big enough to give you meaningful results (especially when dataset is broken down into sub-groups).

3.3 If analysis requires data from more than one source it is important to make sure that these are compatible with one another. For example the data should cover the same time period.

3.4 Is the data source conducive to numerical output (e.g. how should responses to questions in surveys be presented?)

3.5 If the current data collection process is not conductive to the type of analysis planned then it is important to consult with your data provider (see Steps 4 and 5) to discuss the possibility of collecting suitable data that meet your needs.

4. Checking changes over time

4.1 Whether producing a time-series or not checking changes over time to assess whether the most recent data you have is credible when compared to previous data should be undertaken.

4.2 Graphs help to give a feel for data by highlighting trends, and deviations from trends, with far less effort than comparing sets of figures in a table and avoids mistakes in mental arithmetic. This will help to spot discrepancies, whether these indicate an error or not:

- you needn't graph all of your data but it will certainly be useful to graph the main data from each output to spot any obvious discrepancies
- it may be best to plot only one item per time-series for quality assurance to avoid patterns or discrepancies being hidden by values of a much higher or lower magnitude that could skew the axis.

4.3 While graphical displays will help to identify trends in time-series and highlight discrepancies it is not usually practical to produce a graph of all your data as a time-series, but you should check all your data against previous figures where possible.

- Set a threshold for the magnitude of change considered reasonable – this might be based on a statistical threshold or on an understanding of the context which means certain figures are unlikely.
- Calculate both the absolute change and the percentage change between intervals and compare the resulting values to your thresholds.
- Consider which packages will help with this kind of analysis. For example, Excel makes this easy through the use of conditional formatting, but Access and SAS can be helpful as well. This will help to identify anomalies at finer levels than an overall total, e.g. errors due to miscoding.

4.4 If you are producing a time-series consider whether it is reasonable to construct a time-series of the data you have.

- Is there enough data? Constructing a time-series with only two or three points is often inappropriate as it could give an impression of a trend that may be misleading.
- Are you comparing data from regular intervals? Although you may wish to compare data that is not from regular intervals you should consider if there are any cyclical variations that could impact on the interpretation of the results.

- Have there been major changes that might affect the comparability of historical data and more recent data? While some policy or sectoral changes over years can often be expected, reclassification of variables and definitional changes may affect the validity of comparisons.

4.5 When publishing time-series you should produce consistent historical data where possible. Where time-series are revised or changes made to methods or coverage you should consider whether these changes make comparisons with historical data unreliable or unsuitable. If this is the case you should consider the following:

- are the figures used in government targets or indicators? If so it is normal to continue measuring on the old basis if possible (if not possible you will need an agreement from ministers before making changes to measurement)
- where it is not possible to produce a historical series on the new basis, you must produce your best estimate of at least the previous year's statistics on the new basis, or the new year's figure on the old basis to allow change to be measured on a like for like basis. Ideally joint running of the system would then allow measurement of the change in definition or process to be quantified. In terms of presentation of the step change this may take the form of an overlap on a graph where there old method stops and the new method starts
- inform users of revisions through a revisions policy for scheduled revisions or, for unscheduled revisions, a statement within your product explaining the nature and extent of the revision and how the figures should be interpreted.

4.6 If something seems unusual discuss it with data providers and/or policy colleagues – make use of their knowledge of changes in policy or within sectors or of the collection process that may explain unexpected data. It is often important to contact the origins of the data, e.g. institutions that have submitted administrative data to a central collection and to ensure that explanations for changes are well evidenced. It's important to be able to explain anything that looks anomalous; if you can't how do you know it's reliable?

5. Explaining changes and discontinuities

5.1 Changes and discontinuities in data (e.g. in the form of a step change in a time series) can occur for a number of reasons. It is important however, that these are investigated and that you are able to explain the reasons for them.

5.2 Reasons for discontinuities may include:

- change in methodology, such as sampling methodology and size, or the method of data analysis used. In the case of surveys, a new contractor may have been used
- changes in question wording, or a change in interpretation of a question on the part of the respondent
- errors in collating and analysing the data

- underlying changes related to the data (e.g. change in policy including those impacting on the priority and political importance of the data; economic or environmental changes).

5.3 Ensure that those points in Chapters 1-3 of this document have been covered and, ideally have your data peer reviewed.

5.4 Policy colleagues may be able to provide possible underlying explanations for any discontinuities (e.g. new legislation or funding streams). Ensure that you do this for both 'positive' as well as 'negative' changes, ensuring that the conclusions are fully evidenced and do not rely on assumptions.

5.5 Contact key data providers to get an understanding of their interpretation of any questionnaires/survey/returns that may have affected response in a particular area.

5.6 Principle 4 (Practice 7) of the Code of Practice states that, 'Where time series are revised or changes made to methods or coverage, produce consistent historical data where possible' in order to present a clear picture of change over time. This should be done in accordance with the SSSC's corporate Revisions and Corrections policy. Where it is not possible, the previous year's figure, at least, should be estimated on the basis of the new method.

5.7 Be aware that any one explanation on its own may be spurious so it is important to have an overall picture of your data and the issues associated with it. Also be aware that the data is not context free and some users and providers may have a bias towards a given type of conclusion.

5.8 Ensure that an explanation for changes and discontinuities are communicated in your publication. It may help to consider how the statistics are presented in order that the user is able to draw informed conclusions. Footnotes or introductory/explanatory notes should be used to highlight key sources of discontinuity, however additional commentary can provide further context and perhaps assist in outlining the relative significance of various contributory factors. Far from being a source of error, changes and discontinuities can provide a fuller picture of issues relating to a particular dataset. You may also wish to provide an explanation of any key points in any associated minute and news release.