Alternative data sources: Best practice for deciding fitness for use

Scott Kilbey, Office for National Statistics (September 2021)

1. Abstract

This paper details ongoing work by the Office for National Statistics (ONS) to assess the quality of data sources to better produce statistics for the public good. It addresses how to measure and quantify data quality and how this can be useful in the production of economic statistics. Building on the Total Survey Error (Groves et al, 2010) and Stats NZ guide to reporting on administrative data quality (2017), this paper will discuss the efforts to apply suggested data quality metrics to one key output – the Supply and Use Tables (SUT). This is of importance due to the need to understand any underlying divergence between the three approaches to Gross Domestic Product (GDP), as well as contributing to the ongoing developments and case studies in measuring data quality. To date, ONS has applied quantitative data quality metrics to two major data sources contributing to the SUT and is working to expand this to further data sources as well as provide complimentary analysis to enable users to understand, interpret and apply these, currently experimental, results. This work will provide an evidence base for analysts to review current methods, such as balancing adjustments, and make appropriate decisions based on the quality of the underlying data sources as well as having implications for selecting future data sources.

2. Introduction

There are three approaches to measure Gross Domestic Product (GDP) – Expenditure, Production, and Income – delivered via our Supply and Use Tables (SUT). In economic theory these measures should give the same value of the UK national economy. However, in practice this is impossible as the ONS does not have access to complete or perfect information, and different data sources feed into the three different approaches.

Whilst research revealed that there is no international consensus when it comes to measuring data quality, there is recent work published by Stats NZ that provided a possible application to the problems of quantifying data quality and making survey and administrative sources comparable.

This work aims to improve the information available to the SUT team by providing a more accurate record of sources feeding the SUT and a data quality assessment framework which should be used to assess these – regardless of source type - starting with the most impactful data sources. This will allow the SUT team to evaluate current assumptions made in the balancing process, whilst providing several additional benefits to ESG – such as assisting in decisions around implementing alternative data sources.

3. Methods

3.1 Stats NZ framework

After a period of researching the best practice in measuring data quality for both survey and administrative data sources, ONS settled on applying the guide to reporting on administrative data quality published by Stats NZ in 2020. This section will briefly summarise the framework, which relies on the following phases:

- Metadata phase
- Phase 1 (the data source in isolation)
- Phase 2 (taking variables and objects from source datasets and using them to measure the statistical target concept and population)

The first phase is centred around gathering information that is often only known by the experts who work with the data source. This involves asking targeted questions that reveal paradata (defined as data about the process by which a data source is collected) which can be used in the following steps. This information is recorded qualitatively, providing context for the data quality metrics that calculated in the next phase.

Phase 1 consists of turning this qualitative information into quantifiable metrics. The framework provides 25 numerical metrics indicating sources of error within six key areas. See figure 3.1a.

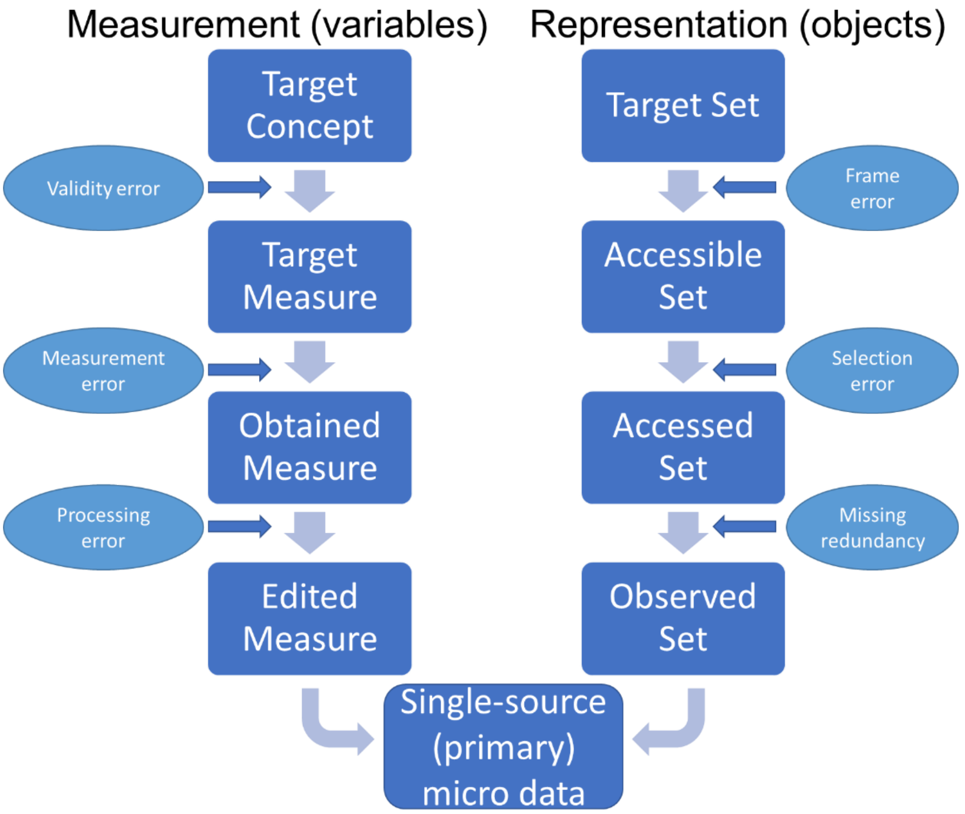


Figure 3.1a: Phase 1 of quality assessment framework, Zhang 2012

Figure 3.1a shows two broad sources of error: the measurement of variables and the selection of objects to be represented by the data. On each side, there are multiple steps away from the desired variable to be measured and the desired object. Each step further away is a chance for error (or deviation from the true reality) to enter the resulting data. There are 25 suggested metrics categorised into these 6 sources of error.

The metrics themselves can be found in this document published on the Stats NZ website, although it is worth noting that not all metrics would be calculated in the same way and not all will be used for every data source.

Phase 2 has a similar approach, with suggested metrics to apply to the derived dataset, however there are no case studies available detailing the use of the suggested metrics and the ONS has yet to complete a case study at this stage.

There is one additional step - Phase 3, dealing with sources of error in production of official outputs – that is set out in a Stats NZ contribution to the Journal of Official Statistics in 2017. However, as the current ONS application of this framework is intended for use in production of the final output it will not be addressed here.

3.2 Applying the framework and additional challenges/analyses

This section details the steps ONS are taking to advance this work for use in the SUT framework. At this stage, much of this is in early experimental and development phases.

The first unanswered issue is how to prioritise and weight the several metrics calculated for a single data source. Even when taking the largest metric from each of the six areas it can be difficult to know which should be given the most attention. There is also interest from stakeholders in whether this can be summarised as a single combined figure.

To address this, ONS have been conducting regression analysis to compare available quality metrics with standard errors traditionally used and produced for survey sources. This line of investigation is based on the hypothesis that certain categories of error may correlate with standard errors – the traditional indicator used to indicate accuracy of survey estimates. Whilst this can only be done for survey sources, it may provide valuable insight into a possible weighting of data quality metrics.

The second area ONS is advancing is the application of this framework to inform methodological decisions. The economic formulae for GDP must balance in theory across the three approaches (expenditure, production, and income), although in practice the data available to build the Supply and Use Tables (SUT) does not give this result. Therefore, balancing adjustments are required to achieve the final published result. The size and location of these adjustments, relative to their position in the SUT matrices indicate a level of trust that ONS has in different data, informed by multiple sources. It is possible to display these adjustments in a heat and tree map.

It is also possible to produce the same tree map and replace the balancing adjustments data with the data quality metrics. This allows analysts and methodologists to have a visual indicator as to whether the adjustments being made are more focused in areas with lower quality sources, according to the framework.

A mock-up of the supply side tree map with quality scores indicated is shown in figure 3.2a. The size of each section represents the monetary contribution of a source to the SUT with the colour indicating quality. At this stage, many of the values are placeholders as research is ongoing.

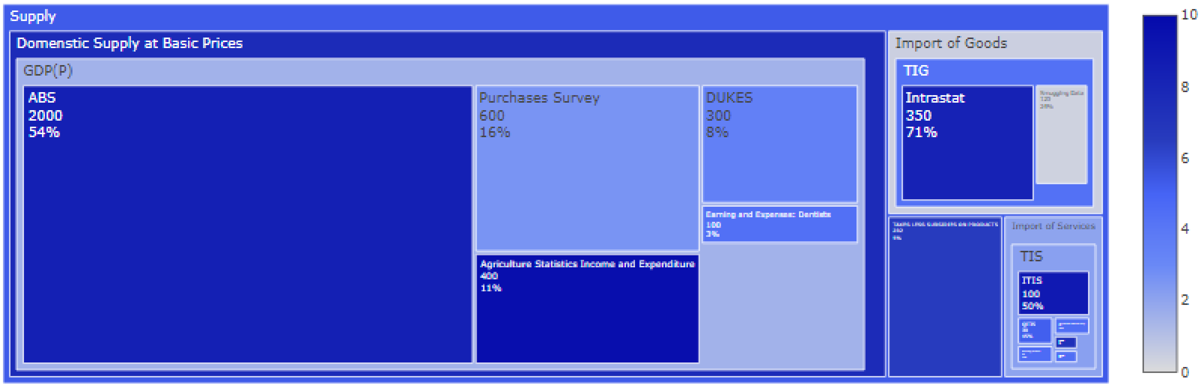


Figure 3.2a: Mock-up of the data quality dashboard in development at ONS, a tree map of sources on the supply side of SUT with colour scale to indicate data quality

Finally, ONS plans to integrate this framework into the acquisition of new sources. Applying this framework to both existing data sources and possible alternate data sources will provide a quantitative evidence base to inform discussions amongst subject specialists. Data source confrontations consider the strengths and weaknesses of different options with which to produce National Statistics. Whilst this should ultimately come down to a decision between relevant experts, this work seeks to provide additional context and inform decisions, rather than make them.

4. Results

As published, Stats NZ have applied this framework to a few select case studies. ONS is working to apply this framework within the Supply and Use Table (SUT) framework – one of the key Economic Statistics publications. In this early stage, this has been applied to two data sources:

- Annual Business Survey (ABS) – Survey Source, internally run
- Pay as you earn (PAYE) tax data by industry – Administrative, external

This paper includes the trends found so far in the ABS (service sectors only) (figure 4a) and the PAYE data set (figure 4b). The decision to display trends is to focus on and demonstrate the viability of the data quality framework rather than the merits of individual data sources.

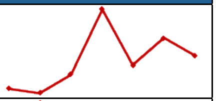
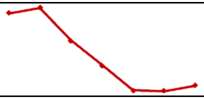
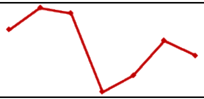
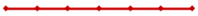
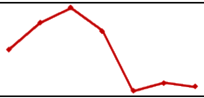
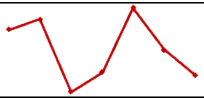
Source of Error: Category	Data Quality Metric	Description	Trend
Validity	Percent of inconsistent records	Errors identified that can't be reconciled	
Measuremer	Percent of units which fail checks	Percentage of items that fail automated checks	
Procesing	Modification rate	Percentage of returns manually cleared after identified errors - no indication of changes	
Frame	Overcoverage	Zero tolerance for duplicates in 'batch tests'	
Selection	Adherance to reporting period	Percentage of returns identified as unacceptable reporting period	
Missing/ redunacy	Unit non-response rate	Factored into planning	

Figure 4a, ABS data quality metrics for service sectors only. Batch tests refer to the first round of basic checks applied to returns, including identification of duplicates.

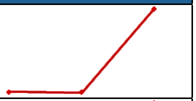
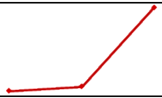
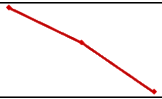
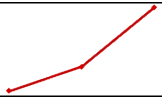
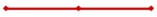
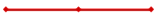
Source of Error: Category	Data Quality Metric	Description	Trend
Validity	Percent of items affected by respondent comprehension of questions asked	Average absolute revision across known IOG	
Measuremer	Item non-response	IOG unknown - measured as a percentage of pay variable (numbers unknown)	
Procesing	Percent if transcript errors	Decrease in RSE(Relative Standard Error) when aggregating IOG to Section.	
Frame	Undercoverage	Undercoverage due to fraud and illegal activities	
Selection	Adherance to reporting period	RTI data, so all data adheres to reporting period	
Missing/ redunacy	Pecenatge of duplicate records	Assumed rare, close to zero	

Figure 4b, PAYE data quality metrics. RTI = Real time indicator, IOG = Industry Output Group, a method to assign businesses to standard industries.

Each table displays the six source of error categories, the largest available metric calculated, a brief description of what was calculated, and the trend observed. For both data sources the metrics cover up to the 2018 annual data, however there was more historical information available for the ABS, hence the longer time series.

From this work a few conclusions can be drawn. Most importantly, it is possible to observe aspects of data quality over time and see them improve or become more severe. Secondly, it was more difficult to obtain additional information for the external data source. However, once the process of collection, revisions and processing was understood it is possible to quantify several metrics.

These early case studies also indicate that there was more variation for the ABS (survey) than the PAYE (administrative). This was an expected result, but more sources need to be analysed before any conclusion can be drawn here.

Regarding the question around weighting of the six metrics, it is possible that analysts may only be interested in specific areas depending on how the data source is used. For example, a benchmark figure or pattern series. However, for the ABS it has been possible to observe how the metrics correlate to standard errors produced in processing the survey results. Figure 4c shows the relative standard error (RSE) - that is standard error as a proportion of the survey estimate - of employment costs for service industries superimposed with the validity metric (metric 3) – number of errors identified that could not be reconciled.

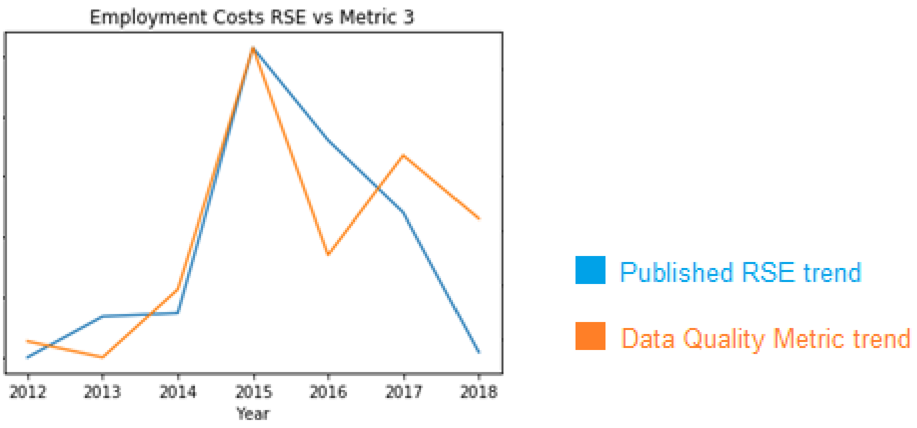


Figure 4c, Metric 3: Validity – Percentage of inconsistent records

In this graph the blue line represents the size of the relative standard error for employment costs. In orange is the percentage of returns containing irreconcilable errors. While not a perfect match, there is some similarity in growth rates, of note the spike in 2015 for both time series.

For transparency, other available metrics did not show as strong correlations. In one case a reversed pattern was evident – though this metric represented the percentage of survey returns that went through a manual clearance process after an error was identified. This may indicate that the manual clearance process is inversely related to the standard error. See figure 4d.

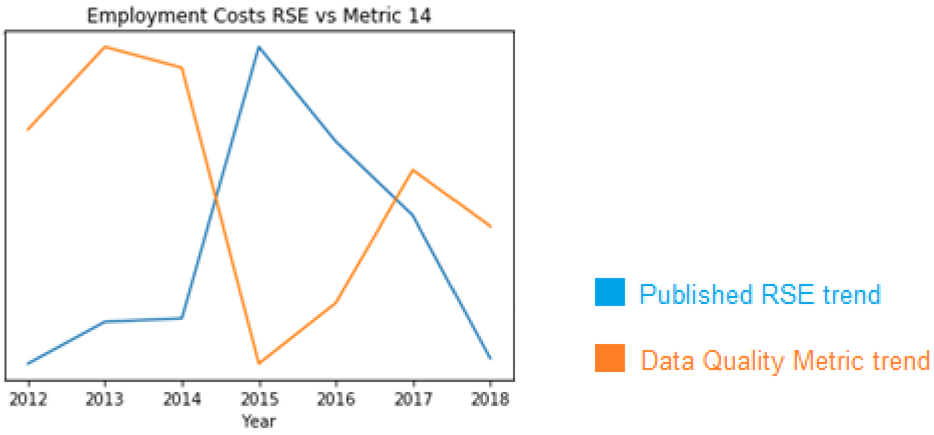


Figure 4d. Metric 14: Processing – Modification rate

In this graph the blue line represents the size of the relative standard error for employment costs. In orange is the percentage of returns manually cleared after errors were identified. These time series show an opposite growth rate pattern for most of the time period.

It is entirely possible that these are coincidences – correlation does not necessarily imply causation. However, these early results indicate the additional analyses that can be carried out when researching survey sources. Before any decisions can be made around the weighting of data quality metrics more (survey) sources need to be analysed to determine whether this line of investigation is viable.

5. Summary

Even though this work is at an early stage, the output so far is promising. It has been established that trends in aspects effecting data quality can be quantified and observed to change over time. It is also clear that results can be delivered iteratively to allow analysts to consider the information available and direct the project to better meet their needs.

This work will provide an evidence base for analysts to review current methods, such as balancing adjustments, and make appropriate methods changes based on the quality of the underlying data sources. In the long run this framework for data quality can provide quantitative information to aid making decisions around the suitability of alternate data source in producing the UK National Accounts.

6. References

- Total Survey Error (Groves et al, 2010)
- Guide to reporting on administrative data quality, Stats NZ (2020)
    - Including Quality indicators for Phase 1 errors
- Topics of statistical theory for register-based statistics, Zhang, Li-Chun, (2012)
- Extending TSE to Administrative Data: A Quality Framework and Case Studies, Stats NZ (2017)