# Quality Changes when Moving from a Survey to an Administrative Source

Daisy Hamer

Office for National Statistics, Newport, United Kingdom; daisy.hamer@ons.gov.uk

**Abstract**

# National Statistical Institutes are increasingly looking to use administrative data for statistical purposes, often replacing survey questions or whole surveys with such data. The purpose of this research is to look at possible changes in quality that result from moving from a survey to an administrative source. The first step to carrying out this piece of research was to conduct a literature review looking at what previous investigations have found and see whether any practical advice could be drawn from this. The second step was to find case studies, working with areas at the Office for National Statistics (ONS) which are, have, or are considering replacing survey data with administrative data and investigating quality changes that result. Research is still ongoing but current findings point to some benefits and some issues that arise as a result of replacing survey with administrative data. These are areas which need to be considered in the development of guidance for statistical producers. This paper will summarise findings from the literature review and the research conducted so far.

Keywords:administrative data, quality.

## Introduction

### **1.1. Rationale and Approach**

National Statistical Institutes (NSIs) are increasingly looking to use administrative data for statistical purposes, often replacing survey questions (supplementing) or whole surveys with such data. It is important to understand the potential benefits and issues with introducing and using administrative data, therefore the Quality Centre (QC) at the Office for National Statistics (ONS) carried out an investigation to assess changes in quality resulting from the use of administrative data to supplement or replace survey data. The first step in carrying out this research was to conduct a literature review looking at previous investigations, from ONS and other NSIs, to understand the approach taken by others and see whether any practical advice could be drawn from this. The second step was to find case studies, working with areas at ONS which are, have, or are considering supplementing or replacing survey data with administrative data, and investigating quality changes that result from this move. This research is undertaken with a view to developing guidance.

## Literature Review

**2.1 Introduction**

The investigation began with assessing reports of major European studies such as ESSnet projects. These reports were also used to find other previous reports that could be assessed and to find contacts. Various NSIs were contacted directly and asked about what work they had done in areas of quality and administrative data. Useful information was assessed and added to the review.

This section summarises the key findings of the literature review. It discusses what has been found in both:

* how quality in administrative data is assessed and
* how quality changes when administrative data is introduced to supplement or replace survey data.

**2.2 Assessing Quality**

This section summarises what the review found on methods of assessing and measuring quality.

The review found that there are various different ways of measuring quality in administrative data. The majority of the reports based their methods of assessing quality on the ESS quality dimensions but removed certain dimensions or included new ones. An example of this is a Blue-ETS project (2011) excluded the ESS dimension “Accessibility and Clarity” from its method of administrative data quality assessment but included a new dimension called “Technical Checks”, a dimension that is exclusive to administrative data and looks at how useable files and data within the files are. This added dimension seems useful and should be something that NSIs consider and report on when discussing the quality of administrative data sources.

Some of the reports also broke down quality dimensions into composite indicators. Composite indictors seem like a logical way of assessing quality, however an ESSnet report (2011) states that it can be difficult to derive a composite indicator that is meaningful, and therefore recommends that quality should be assessed in no further detail than acceptable and unacceptable. The report also states that composite indicators are only useful for looking at quality improvements of one output over time and not for comparing different outputs. This indicates that it would not be a useful method for comparing survey and administrative data. Along with this there is subjectivity involved in setting reference values which could cause inconsistency in a quality assessment.

The review also noted that a possible way of measuring accuracy in administrative data would be to develop a framework similar to Total Survey Error (TSE). Such a framework looks at errors which come from every stage of the survey process and is calculated using the Mean Squared Error formulation. It seems that some survey errors may be relatable to errors in administrative data. For example, in specification error, rather than the issue being a survey question implying a concept different to what is being measured, it could be that a definition used in an administrative dataset could be different to what needs to be measured for the output. This method of measuring quality seems like it may be a potentially useful tool for comparing survey and administrative data as the total error of each could be compared against each other, although it is difficult to say how meaningful this comparison would be.

Overall there are many ways of assessing quality in administrative data and it is difficult to say whether there is a best method. It should also be noted that many of the methods of assessing quality reviewed look solely at how quality can be assessed in administrative data and not how to compare the quality of survey and administrative data.

**2.3 Quality in Administrative Data**

This section looks at what the review found from previous literature available that that looked at quality benefits and issues that are common in administrative data or that arise when introduced to supplement or replace survey data.

The main issue highlighted in the review was accuracy, which can be negatively affected by problems such as under- or over-coverage, or incomplete or inconsistent data. Other problems can be with data collection or with the ways in which the data are compiled. There may be no quality checks on the data, there could be biases in the way data is input, or data may not be recorded appropriately or at all.

An example of accuracy issues comes from a report by Statistics New Zealand (2014) that looked at using administrative data to determine population statistics in their census. The report highlighted that the administrative data used had accuracy problems in that there was both under- and over-coverage. The report did find that the more administrative data sources that were linked together, the more accurate the data seemed to be, although there were still issues with coverage of certain populations. Because of this it was decided that at that time it was not feasible to use administrative data for population estimates.

Another quality issue that was highlighted often in the review was that there can be problems with timeliness in administrative data. A report by the Australian Bureau of Statistics (Pink, 2011) stated that there is often a delay between when the data are collected and when they are available for use and there may also be delays in how long it takes for data to be collected. This is not only an issue with timeliness but can cause a trade-off between timeliness and accuracy; more timely data will be less complete and therefore less accurate, and waiting for more complete data would make the data less timely.

Trade-offs are not uncommon when considering quality and there can be trade-offs between quality dimensions when administrative data is introduced. The quality dimensions are not independent of each other and the introduction of administrative data can cause an improvement in quality in one of the dimensions and a worsening in another.

The review showed that accuracy and timeliness were found to be common quality issues for administrative data in various types of statistics including business and census statistics. There are ways of improving the quality in these areas but issues can still occur and NSIs should consider the trade-offs between the dimensions.

**2.4 Discussion**

The reports discussed look at various different quality issues in relation to administrative data. One of the main issues is with accuracy. However, the reports also show that there are some methods of dealing with the issues that arise, such as linking data sets (although this can be prone to error).

This review also looked at ways in which quality can be measured in administrative data. It found there are multiple ways of doing this, and each report reviewed measured quality in different ways. The most appropriate way of measuring quality may differ depending on how and where administrative data are being used.

An issue with compiling the literature review has been is that there are a few examples which look at quality changes when survey data is supplemented or replaced with administrative data. As a result of this, the Quality Centre collaborated with other areas in ONS working on a project to measure the quality changes resulting from the introduction of administrative data.

**3. Case Studies**

One of the case studies relates to International Migration Statistics.

Long term international migration is currently calculated using the International Passenger Survey (IPS) but in 2014 ONS considered the possibility of supplementing or replacing this with e-borders - known as Semaphore – data (ONS, 2014). This is data collected by the Border Force using the Semaphore system, which is a system which collects the travel information of all individuals travelling into and out of the UK including the individuals’ passport and booking data.

ONS produced a report assessing the possibility of using Semaphore data and found that there are both quality benefits and issues. Similarly to the literature review, most of the quality issues were related to the “Accuracy & Reliability” and “Timeliness & Punctuality” quality dimensions.

The report found that the Semaphore data has positive and negative aspects for “Accuracy & Reliability”.

One negative was that over-coverage was found to be a particular issue in the data as the way the Semaphore system works means that there are often replicate or duplicate records. These can be reduced using linking and other de-duplication strategies but there are still issues that need to be resolved and the report stated that this is an issue that will not be one that is easy to resolve. A positive for “Accuracy & Reliability” was that Semaphore data has a fairly high level of completeness at 96% which seems like an improvement on the IPS overall response rate of 79% (ONS, 2015).

Timeliness issues with Semaphore data were also highlighted. The IPS has a question which asks whether someone is a long term migrant, but with Semaphore data individuals entering or leaving the country have to be assessed for 12 months before it could be gathered whether they are a long term migrant. This would mean a lag of over 12 months between the end of the reference period and the release of estimates whereas the current statistics have a 5 month lag.

This is another example of a trade-off between accuracy and timeliness. If Semaphore data is used then it has the potential to be more accurate as it would have information on whether an individual is actually a long-term migrant, rather than with the IPS where they just know the individual’s intentions. However, waiting to get this potentially more accurate data will make it less timely, and, because of this, potentially less relevant.

Based on this information ONS concluded that Semaphore data is not currently appropriate for use in Long Term International Migration statistics. However, it was noted that the data is continuously being updated and improved and so it may be appropriate for use in the future.

**3.1. Discussion**

The case studies assessed showed that there are both positive and negative aspects, with regards to quality, to using administrative data to supplement or replace survey data. They also showed that each administrative data source has different quality positives and negatives; there does not seem to be any consistent quality issues or benefits for all administrative data sources or outputs. For example some reports stated that coverage is better in the administrative data source that in the survey counterpart whereas others stated that the administrative data source has serious coverage issues.

**4. Conclusion**

Both the literature review and the case study show that there are many quality issues that can arise from the use of administrative data and that there needs to be continued investigation into these sources, both how and what is measured and how these issues can be solved.

Whatcan also be taken from this investigation is that there are certain quality dimensions and indicators which should be considered before the decision is made on whether to use administrative data to supplement or replace survey data.

Issues that were mentioned most frequently in the case studies were coverage – both under- and over-coverage, timeliness issues such as delays in updating administrative databases, and issues of comparability of definitions used for statistical purposes and those for administrative purposes, or whether a new administrative data based output is comparable to the survey data based output that was previously produced and similar outputs produced internationally.

The quality dimensions mentioned in this investigation were “Accuracy and Reliability”, “Timeliness and Punctuality”, and “Coherence and Comparability” and so it seems the affect on these dimensions are the ones that are most important to consider when introducing an administrative data source to supplement or replace survey data.

It is important to note that just because administrative data has quality issues; it does not mean that it should not be used in statistics, survey data also comes with many quality issues. What should be taken from this investigation is that for both survey and administrative data, whilst improving quality should be a main concern, NSIs should also know what quality issues their statistics have and be able to report this in a clear manner. Therefore, based on this research, the Quality Centre is in the process of developing guidance for measuring and reporting on quality when administrative data is used to supplement or replace survey data.

**5. References**

BLUE-Enterprise and Trade Statistics. (2011) *Deliverable 4.1: List of quality groups and indicators identified for administrative data sources.*

ESSnet on Use of Administrative and Accounts Data in Business Statistics. (2011) *WP6 Quality Indicators when using Administrative Data in Statistical Outputs.*

Office for National Statistics. (2015). *International Passenger Survey: Quality Information in Relation to Migration Flows (Background Note).* November 2015.

Office for National Statistics. (2014). *Semaphore Research Update.* 21 November 2014.

Pink, B. (2011) *Information Paper: Quality Management of Statistical Outputs Produced From Administrative Data,* March 2011, cat. no. 1522.0, ABS, Canberra

Statistics New Zealand. (2014) *An overview of progress on the potential use of administrative data for census information in New Zealand: Census Transformation programme.*