

# Methodological challenges in integrating data collections in business statistics

Paul Smith<sup>1</sup>

<sup>1</sup>Office for National Statistics, e-mail: [paul.smith@ons.gov.uk](mailto:paul.smith@ons.gov.uk)

## Abstract

This paper reviews the opportunities and challenges presented by the integration of data from a range of different sources for the production of business statistics in National Statistical Institutes. It addresses some of the issues such as measuring the quality of inputs and outputs for multiple-source data, the challenge of identifying the right reference times and of disaggregating series of information in time. We also look at some of the methods available for using and dealing with multiple data sources such as dual-frame estimation, benchmarking and balancing, data confrontation, and modelling.

**Keywords:** Administrative data, model-based estimation, quality measures

## 1. Introduction

Several factors are driving producers of official statistics to do more with less – user demands are ever increasing, the resources available for taking surveys are always under pressure, and businesses and individuals are becoming less tolerant of surveys. These pressures are causing efficiency to be sought wherever possible in the production process, and one of the ways in which this is happening is to be more creative about using different sources of information, and putting them together to increase our ability to produce timely information with good quality attributes. At the same time there is a general joining up of data sources, and a desire to exploit data collected from non-survey sources to improve the way in which surveys are constructed.

In this paper we consider business statistics in particular, and review the opportunities and methodological challenges presented by the integration of data from a range of different sources for the production of business statistics in National Statistical Institutes. It addresses some of the issues such as measuring the quality of inputs and outputs for multiple-source data, the challenge of identifying the right reference times and of disaggregating series of information in time. We also look at some of the methods available for using and dealing with multiple data sources such as dual-frame estimation, benchmarking and balancing, data confrontation, and modelling.

## 2. Data quality

### 2.1 Input data quality

The first issue to consider is the quality of the available data. Quality in surveys has been an increasingly important topic over the last decade, and in Europe has followed on from work to codify the dimensions of quality. Now a regulation on quality means that considerable efforts have been made to calculate metadata about quality to support the series being produced. Metadata for this purpose come in a variety of styles; in some cases they are readily calculated through theory, such as sampling errors. In other cases the quality aspect which is of primary interest is not easily measured, such as non-response bias, but a relatively easy calculation – the response rate – gives an indicator for the magnitude of either the bias, or perhaps the risk that the bias will be large enough to affect the interpretation of the statistics. In a few cases, such as measurement error, there is very little that can be done with survey data and the only real way to measure the quality is to do an expensive follow-up study. In other dimensions there is no direct quantification (for example, relevance), and then only circumstantial information can be provided.

In purpose-designed surveys the concepts and definitions are generally those required for the statistical outputs. There may be some adjustments between the target outputs and what is actually collected, for example to make it more practical for businesses to provide the data (for example by asking them about concepts which they already record for accounting purposes), and possibly for consistency between surveys (although more usually survey concepts will be harmonized so that there is no need to adjust questions for differing concepts). The surveys are also designed according to certain accuracy criteria, principally through survey design methods such as sampling.

Quality for administrative data is normally achieved through *processes*. The target will be some statutory or regulatory requirement, and businesses will need to collect relevant data in order to comply. Therefore there is no adjustment of the concept, and in many cases little concession to whether the information will be easy to provide. Where the administrative concept coincides with the statistical one it is likely that statistics are *already* derived from the administrative system, and therefore that the administrative source is the only one required. Only the issue of timeliness may engender the need for a survey. For example, in the UK statistics on the profits of companies are available from their accounts, which are filed through Companies House. However, there is a long lag before full accounting information is available, so there is a statistical survey collecting information on profits, which provides information up to a year sooner than the administrative source. It is, however, subject to additional uncertainties such as sampling error.

One of the main *measurable* accuracy components, sampling variability, becomes zero in administrative data sources, with the result that the balance of quality is quite different. Other accuracy components therefore become more important, and may be strongly affected by the processes behind the collection of the data. Different variables in the administrative data can have substantially different quality attributes because of their administrative use and the procedures associated with their collection. One long-standing example in the UK was the collection of classification information as part of the administration of Value Added Tax. This was a small and relatively

unimportant component for the tax department, whose main target was to collect the correct amount of tax. But the same data are used as one source for the UK's business register, and this meant that the classification information was very important for statistical purposes. Considerable negotiation was needed to increase the quality checking of classifications through changes to the quality assurance procedures in the tax office.

## **2.2 Output data quality**

Where surveys are used as the only source for outputs, the output quality is generally well-defined and the metadata definitions and methods are well-known. Equally where statistics are derived directly from an administrative source, the quality components will be at least partly well-known.

Using several sources of information in a single output presents challenges in determining the quality of the combined outputs. The methods in use will often be more complex and therefore the quality measurement methods will also be more involved. Most of these concerns will be discussed in more detail under particular methods below, but it is worthwhile considering the relation between data sources which may lead to a combination being a useful strategy. In this context there is a strong case to consider cost as a major factor – although it is not a quality component in the European system, tradeoffs between quality and cost are critical.

The quality and coverage may be substantially different in two sources. One may have broad coverage but poor accuracy, while another had restricted coverage but good accuracy. Typically good accuracy will be associated with high cost, so combining sources provides an opportunity to adjust the poor accuracy, low-cost source with wide coverage to be in line with the measures in the good accuracy high-cost source.

A similar adjustment may be made in the case where the administrative source measures a concept which is not exactly that required for the statistical source, in which case a small survey collecting the statistical concept may provide a suitable adjustment from a comprehensive source measuring the administrative concept.

## **3. Combinations of sources**

Many methods have been developed for using two sources of data together. Indeed, many of them are so familiar that they no longer seem to be particularly challenging. A range of methods is listed here with some brief discussion of their uses, and some comments on their impacts on output quality.

### **3.1 Frame and sample information**

Most business surveys in the developed world are based on frames which contain detailed information about the business population, usually derived from administrative sources. This information is used in sampling, and then again during validation and editing, and in survey estimation. Without this information business surveys must use methods with considerably poorer quality (for example area sampling). Quality measures for sampling from fixed populations (and for estimating

using auxiliary information, see 3.4) are well known, and further measures of coverage are often available.

### **3.2 Dual-frame surveys**

There is sometimes more than one source of administrative information covering the population, or a substantial proportion of it. In these cases it is possible to sample from both frames, and then to combine the sample survey information with information from both frames.

In a restricted fashion, the UK's business register is like a dual-frame survey, because it is derived from two matched sources, VAT and Pay-As-You-Earn income tax (PAYE). In fact procedures are in place to ensure that the register's size is not inflated through the presence of unmatched units, and therefore one part of the dual-frame population remains unestimated. Nevertheless this gives good circumstantial evidence of the quality of the coverage of the two sources.

Dual frame surveys depend on the probability of a sampled business responding being independent of their chances of appearing on one or other of the frames (perhaps conditional on some known information). Then the probability that the business would appear in the sample from one of the frame sources can be calculated and used in an appropriate weighting procedure. There is a Pension funds survey in the UK which has historically used two frame sources, but this is particularly complicated by the two frames having a different unit – one covers pensions schemes, the other pension funds.

### **3.3 Multiple surveys, benchmarking**

Where several surveys cover the same population they can be used together. In the UK there are several examples where detailed data are collected on some surveys and used to provide a breakdown from another survey of the same population at a different periodicity. In its simplest form a small (and therefore variable) monthly survey may be benchmarked to a large (and therefore more accurate) annual survey. Breakdowns of capital expenditure by contrast are collected quarterly by the ONS and the proportions are used to estimate the annual breakdowns from an annual survey where only totals of capital expenditure are collected.

In these cases the quality measures are typically complicated to derive. Using one survey's results to estimate something in another ideally requires that the sampling variability in both surveys is accounted for, whereas it is often assumed that the first answer is known exactly before application in the second survey. Further, the impacts of non-response or measurement error in the first survey on the outputs from the second may be particularly challenging to derive.

### **3.4 Auxiliary information in processing**

Even in cases where the administrative data concept is too far from the statistical concept to be used directly as the output variable in published statistics, it may have several uses during the statistical processing. If the administrative data variable is related to the variable(s) being measured in the survey, then it will be useful in the

sample design stage, and also in (model-assisted) estimation (Särndal, Swensson & Wretman 1992), and this is a common use for administrative and historic data on business registers. In this case it is not even a requirement for the data to be correct, only for it to be correlated with what is being measured. The better the correlation, the better the accuracy (in variance terms) of the final outputs.

Auxiliary information may also be useful in data validation and editing, although such uses are much less common. One area where this is being considered in the UK is the use of statutory VAT returns as auxiliary information in the validation process for business surveys. There have always been a few validation tests which are based on register data, but the use of up-to-date information from the administration of VAT which has not yet been included on the business register has the potential to make savings in the amount of follow-up with businesses and the possibility to catch errors that would otherwise have escaped detection.

### **3.5 Data confrontation**

Where there is more than one source a further possibility is to look for instances where the information from the two sources is most discrepant, a technique which has been called *data confrontation*. In these instances there is a likelihood that one or both sources contains an error or anomaly which would be worthy of follow-up investigation.

Another type of data confrontation is balancing, a process mostly encountered in the construction of National Accounts, where sums of components must come to totals measured from different sources, in two or more dimensions. This has many similarities with calibration, where one total is made consistent with a second variable measured two ways. In balancing the process is usually informed by additional information on the relative accuracy of the different components, and also takes account of known extraneous factors which typically cannot be modelled.

## **4. Mode effects**

The challenge of combining data from different modes is becoming more pressing, since there are savings to be made through the use of new technologies for data collection, but also issues because it is unclear whether there are differences between the data collected by different modes. We can divide this problem into stages; first, to assess whether there is a difference, and second, if a difference is found, to use the information collected through the two modes in the best way possible to produce estimates.

### **4.1 Detecting mode effects**

The introduction of a new survey mode is an opportunity to instigate an experiment within the survey, which can be used to measure the difference in responses between the two modes (see Van den Brakel & Renssen 1998, 2005 for methods for designing and analysing experiments in the context of sample surveys). Part of the mode effect may come from the contact probability – for example, it may be easier to ignore a letter rather than a telephone approach, or it may be that the person filling in the

questionnaire has to gather quite a lot of information before responding, so that it must be written down anyway. So the differences may manifest themselves through a difference in the response pattern, or through a difference in responses after the response pattern is accounted for.

Note that the lack of a detectable mode effect is not an immediate reassurance that such an effect does not exist. There are examples where it is concluded that because no significant effect has been detected, there is no effect, but where the power has not been considered. If the power is low, there will be almost no chance that a change will be sufficiently large to be detected.

#### **4.2 Dealing with mode effects**

There is much less experience with how to deal with mode effects in making survey estimates. If a difference is detected between two modes, then it will be necessary to make a judgement about which is most closely measuring the concept of interest, and therefore which is the one to which we want to adjust.

In the case of a change in the response patterns, it may be possible to model response based on some characteristics of the business, and in this case it will be possible to make a weighting adjustment. Equally, it may be that there are no good predictors amongst the known variables on the frame, in which case there will be no weighting adjustment possible, and we return to trying to assess the difference using the difference in experimental treatments appropriately estimated. A further, though expensive, alternative might be to undertake a non-response follow-up study to gather information on the characteristics of businesses which failed to respond.

Where the numbers which are delivered differ between different modes, the only approach to putting them together will be one based on a model allowing for concepts to be measured in different ways (perhaps a measurement error model). Assessing the difference between the modes using the difference in experimental treatments will be an important input to such an approach, though appropriate estimation may use known auxiliary variables to increase the power in the estimation of the difference.

### **5. Temporal adjustment**

One of the challenges which has already been addressed in several NSIs is how to move from an administrative dataset with one periodicity to a statistical dataset with a different one. If the statistical period is longer this is trivially straightforward, but generally the statistical requirement is for shorter periods, and this means temporal disaggregation of the administrative series. Let us for simplicity take the administrative data as yearly and the statistical data as monthly. Temporal disaggregation can be undertaken if there is some information on the within-period pattern of movement which can be used to infer the in-year pattern. This could come from different year-ends within the annual data, or from monthly data on a subset which can be regarded as sufficiently representative of the within-year changes. For a detailed treatment see Dagum & Cholette (2006).

In the UK VAT responses are required at different times and on different periods for different businesses, with the largest reporting monthly and the smallest annually (and in different months spread across the year). This provides some good information for disaggregation, but is also quite a complex model to describe and fit. It is not very clear how well such a model will respond to rapid changes in economic conditions.

## **6. Modelling**

Full flexibility in combining different datasets can be obtained through a modelling (or model-based estimation) approach, and there are many possible techniques for modelling which can be employed according to the available data sources and their characteristics. With a modelling strategy the estimates may be sensitive to the choice of model, or to the way in which it is used (for example whether the estimates depend on extrapolation beyond the observed data, or merely estimation of missing points within the data). One of the goals of a good model will be that it is relatively insensitive to the model assumptions, and will therefore remain valid under reasonably broad conditions. Where the models fit well and the data are collected according to good designs, the outputs from these approaches can have good quality characteristics. A classic example from social surveys is small area estimation, typically fitted by some random coefficients model (Rao 2003); extending these models to deal with the greater disproportion in weights found in business surveys is a major challenge (Hidioglou & Smith 2005) which has not yet been satisfactorily solved. Nevertheless modeling has the potential to provide a huge palette of methods and techniques for combining data from different sources.

A particular challenge of modelling for business surveys in the UK is in the recently realised Business Register and Employment Survey, a combination of parts of two previous surveys which involves collecting information for reporting units (enterprises) and for their constituent local units (sites). However, in some cases – particularly for the smaller businesses – the local unit information is not collected directly, but is modelled based on a combination of frame information and information from larger businesses, with an associated increase in the uncertainty but reduction in the cost.

## **7. Discussion**

Almost all combinations of data sources aim to produce more from existing data sources (possibly with the further aim of scaling back the statistical component of data collections). To this end they are about combining data which match imperfectly, and the only real approach to this is to use a model to combine the sources. The choice of model will often be subjective, and it will be important to assess the sensitivity of the output to the model structure and to the fitted parameters. The consideration of the sensitivity will also provide information on the quality of the outputs which can be used to judge how well they are suited to their purposes. Not all of the quality components in the European system can be measured in this way, but many of the ones most useful for comparison with other measures can be made available.

In the case of multiple modes there typically will not be much if any additional data – the main drivers are for usability and low cost. In these cases the target is more likely to be to maintain existing levels of quality, and this will often critically depend on how well the different modes manage to collect comparable information. Again some of the quality components may become harder to measure, but there are analytical techniques which can be used with the new technologies (for example experiments within surveys) to assist in both the transition and the measurement of quality.

## References

- Dagum, E.B., Cholette, P.A. (2006) *Benchmarking, temporal distribution, and reconciliation methods for time series*. Springer, New York
- Hidioglou, M.A., Smith, P. (2005) Developing small area estimates for business surveys at the ONS, *Statistics in Transition*, 7, 527-539
- Rao, J.N.K. (2003) *Small area estimation*, Wiley, Hoboken, New Jersey
- Särndal, C.-E., Swensson, B., Wretman, J. (1992) *Model assisted survey sampling*, Springer, New York
- Van den Brakel, J. A., Renssen, R. H. (1998) Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, 14, 277-295
- Van den Brakel, J. A., Renssen, R. H. (2005) Analysis of experiments embedded in complex sampling designs. *Survey Methodology*, 31, 23-40