

Using Administrative Data to Supplement Statistical Surveys

8.1 Introduction

This chapter presents an overview of different models for using administrative data to supplement data collected in statistical surveys. It shows how a mixed-source approach can be used to produce statistics at lower cost, better quality, or both.

Many of the issues relating to using and linking statistical and administrative data have already been covered in Chapters 4 and 6, so are not repeated here. Instead this chapter focuses on the different models for using data from a mixture of administrative and statistical sources to produce statistical outputs.

8.2 Mixed-source Models

1) *The Split Population Approach*

In this model the statistical population is split into two or more parts for data collection purposes. This approach is very similar to that used for the maintenance of the Australian statistical business register, as described in Chapter 7.3. Data from administrative sources are used for units where these data are of sufficient quality, and statistical sources are used for the remainder of the units.

A typical scenario for a business survey is that data for relatively small businesses with simple structures are taken or derived from tax returns, whereas surveys are used to collect data from the key units (usually those that are largest and/or have the most complex structures). For the section of the population for which tax data are used, the statistical and administrative units are likely to be identical, or very similar, and the impact of the difference between statistical concepts and classifications and their administrative counterparts is likely to be minimal, or at least can be easily modelled.

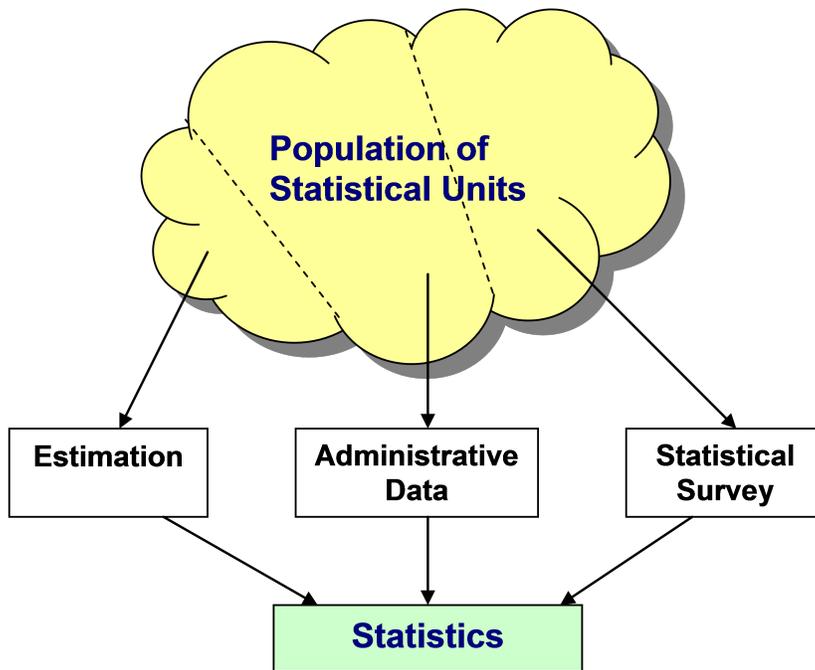
The remainder of the businesses are typically those that have the greatest individual impact on the quality of the statistics, and therefore are the ones for which it is most important to have accurate data. These units are also likely to be the ones with the most complex structures, often requiring profiling (as described in Chapter 4.5) in order to define the correct statistical units for which data are required. These statistical units are often combinations of administrative units, or parts thereof, and whilst some variables such as employment can often simply be summed to give the correct total, others, such as sales and certain other financial variables can not, as they include a certain amount of intra-unit trade, such that a simple summation would result in over-counting.

A practical example of the split population approach in business surveys is the Unified Enterprise Survey conducted by Statistics Canada. This brings together annual business data requirements, combining several previous surveys. Administrative data are used instead of data collected through statistical questionnaires for over half of the enterprises in the survey that have a simple structure, resulting in reductions in the statistical response burden of almost 40%.^[1]

Where the statistical population is people or households, it may be the case that surveys are needed for special groups such as students, migrant workers or those with two or more residencies. These are all potential examples of units for which administrative data may not be sufficiently up to date or accurate, particularly concerning location.

As mentioned several times in previous chapters, consideration must also be given to units not covered by administrative registers, such as illegal immigrants or businesses operating in the informal economy. Statistical surveys are likely to be only of limited use for such groups, so an element of estimation may be needed, thus introducing a third source to be used in the production of the required statistics. This model is illustrated in Figure 8.1 below.

Figure 8.1 – The Split Population Model



2) The Split Data Approach

In this approach, a population of statistical units, and a data requirement are identified, for example the population could be all persons living in a particular country, and the data requirement could be the usual set of variables required for a population census. Instead of providing all of the variables for part of the population, as in the split population model above, under the split data approach, administrative sources are used to provide some of the variables for all of the population (a third approach is also possible where administrative sources provide some of the variables for some of the population).

The split data approach does not, therefore reduce the number of questionnaires or interviews required to collect the data, but does reduce the volume of data to be collected in each questionnaire or interview. It is usually most relevant for large and complex data collections where many variables are required, hence the example of the population census. Administrative and survey data need to be integrated for each individual unit in order to produce the data set used for statistical outputs.

The split data approach is often used during the transition to the sort of register-based statistical system described in the next chapter. Typically, the variables in the statistical data collection are replaced by their equivalents from administrative sources over a number of survey periods. Table 8.1 below illustrates this process showing data sources for the Finnish population and housing census.

Table 8.1 – The Split Data Approach in the Finnish Population and Housing Census 1960-2000

	1960	1970	1980	1990	2000
Demographic Data	Q	Q/R	R/Q	R	R
Economic Data	Q	Q/R	Q/R	R/Q	R/Q
Education Data	Q	Q	R	R	R
Household and Family Data	Q	Q	R	R	R
Dwelling Data	Q	Q	Q	R	R
Business Premises Data	Q	Q	R	R	n/c
Building Data	Q	Q	Q	R	R

A cheaper alternative may be to decide that if data not provided by a particular date, particularly for units that are not vital to the survey results (e.g. smaller businesses in a business survey), they are instead taken or derived from administrative sources. This allows any response chasing resources to be focused on the units that are considered most important, which should mean that any bias from using administrative data rather than survey data is minimized. This can also help to improve the timeliness of the survey results. As with any quality-related issues, a compromise between cost and the different dimensions of quality (see Chapter 5) is inevitable.

Administrative data can also sometimes be used as a basis for imputing missing survey data for linked data files[2].

5) Using Administrative Data for Estimation

When a sample survey is used to collect statistical data, it is often necessary to use estimation techniques, particularly if population totals (rather than proportions) are required. Some basis to estimate the values for the non-sampled part of the population is therefore needed. Sometimes this process can use variables from the survey frame used to draw the sample, but in some cases it may be possible to improve accuracy by using data from administrative sources as auxiliary variables in the estimation process[3]. In practice many examples of this approach concern using administrative data to improve estimates for small areas[4].

8.3 Further Considerations

In any complex statistical processing system using multiple sources, it is vital to consider the role of metadata, particularly those metadata relating to the source of a particular data item. This allows for data items to be treated in different ways throughout the various processes (including unforeseen future processes), according to the way in which they were obtained. Information on the data source is also often a powerful quality indicator, and can help with decisions on the level of quality of statistical outputs.

Using a mixture of statistical and administrative data can be seen either as an end in itself, particularly where the coverage or quality of the administrative data is not seen as sufficiently high to allow statistical data collection to be stopped altogether. It can also be seen as a step in a gradual transition towards a register-based statistical system, as demonstrated in Table 8.1.

Either way, it allows at least some of the benefits of using administrative data to be realised (including cost savings), whilst avoiding some of the disadvantages, such as total reliance on an external supplier and loss of contact with the general public. It gives the possibility to compare statistical and administrative data quality, and allows statisticians to become familiar with using administrative data, and to develop new techniques to improve process quality.

For these reasons, mixed-source approaches are currently much more common than purely register-based statistical systems, however, over time, confidence in administrative data is likely to increase, allowing their use to be expanded and further benefits to be realised. As the balance swings further towards administrative data it will eventually become necessary to consider whether to switch to the sort of register-based model described in the next chapter.

[1] For more information, see the paper "Use of Tax Data in the Unified Enterprise Survey (UES)" by Marie Brodeur of Statistics Canada. http://unstats.un.org/unsd/economic_stat/Moscow_workshop/Canada%20-%20Use%20of%20tax%20data%20in%20the%20UES-E.pdf

[2] For example see the US Census Bureau approach in chapter 3 of the publication: Reengineering the Survey of Income and Program Participation, <http://www.nap.edu/catalog/12715.html>

[3] For example see: The Use of Administrative Data Sources for Lithuanian Annual Data of Earnings, http://home.lu.lv/~pm90015/workshop2006/papers/Workshop2006_22_Slickute_Sestokiene.pdf

[4] For example see: Using Administrative Records for Small Area Estimation in the American Community Survey, <http://www.fcs.gov/99papers/mcf.html>