Using Administrative Data in Statistical Registers

7.1 Introduction

The previous chapters have considered the various issues involved in getting access to administrative data, and ensuring that they are fit for use for statistical purposes. Many of these issues are relevant for the day to day management of a statistical register, but will not be repeated here. Instead, this chapter considers ways in which administrative data can be mobilised for the statistical production process through their integration in statistical registers. It first defines statistical registers, and then looks at different models that have been used to integrate administrative data.

7.2 Defining a Statistical Register

There are various definitions of registers, though often with common themes. One of the more widely used is:

"A register is a written and complete record containing regular entries of items and details on particular set of objects." ***[1]***

Typically a register is some sort of structured list of units, containing a number of attributes for each of those units, and having some sort of regular updating mechanism. In this way, many administrative data files can be considered to be registers, but the results of one-off data collections are not.

It could be argued that where statistics are produced directly from a single administrative source, this source should not be considered to be a register, in the same way that survey, or even census results are not normally considered to be registers. This argument is even stronger when the administrative data are used in the form of aggregates rather than individual unit-level data.

A statistical register is a register that is constructed and maintained for statistical purposes, according to statistical concepts and definitions, and under the control of statisticians. Administrative registers can therefore be used as sources for statistical registers, but the reverse would normally be seen as contradicting the principle of the "one-way flow" of data[2].

A statistical register typically plays the role of a data coordination tool, integrating data from several sources, both statistical and administrative. This may be done by linking records using common identifiers, or by using the sorts of matching techniques described in Chapter 6. It may sometimes be easier to use data from a single source, but in such cases it is often difficult to check the accuracy of that source. When several sources are used and integrated within a statistical register it is possible to have a much better view of the accuracy of the data. Unfortunately the negative side of this is that it becomes necessary to have a strategy for dealing with conflicting data from different sources. However, if variables in statistical registers are stored with source codes and dates, automated algorithms can be used to prioritise sources and resolve most data conflicts.

As well as integrating data from different sources, a statistical register may also provide the possibility to derive new variables. One example is that several countries[3] use data on legal form, economic activity classification and foreign ownership in their statistical business registers to derive the institutional sector[4] used for National Accounts.

Traditionally statistical registers have been used as sampling frames for surveys, but they are increasingly being seen as sources of statistical data in their own right, particularly regarding data for small geographical areas, or small sub-groups of the population. Statistical registers can also provide the basis to link data from different sources over time, allowing longitudinal analysis. This approach has been used in several countries to allow studies of cohorts of people or businesses.

7.3 Models for Creating and Maintaining Statistical Registers using Administrative Data

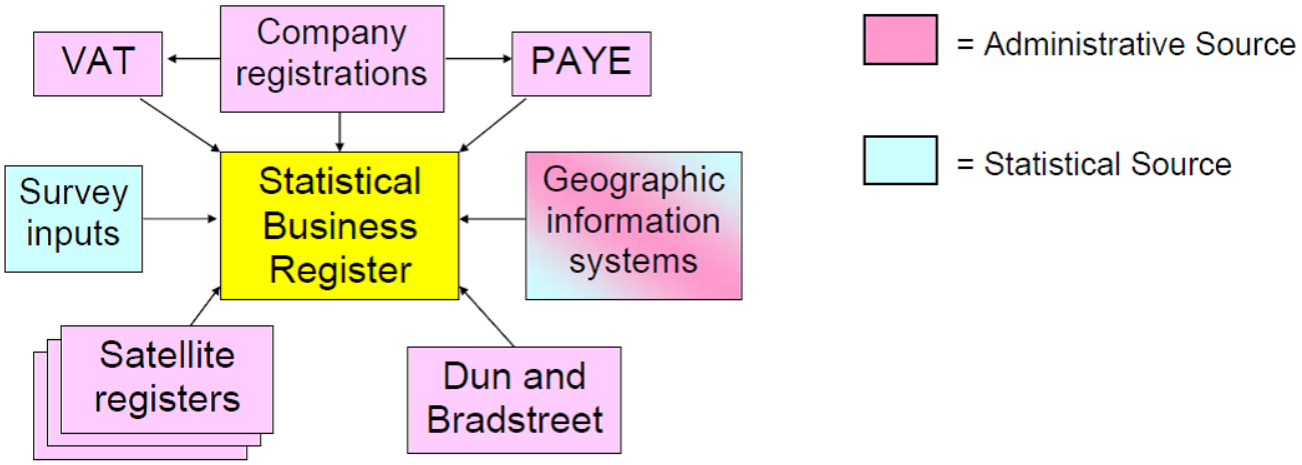
As mentioned above, statistical registers play an important role in coordinating data from different sources. There are many ways in which these sources can be used or combined to produce sampling frames or statistics. This section looks at some of the approaches used in different countries, and for different areas of statistics.

As the sources available differ significantly from one country to another, it is often difficult to export a model, or to define international standards. The different models below should not therefore be seen as recommendations that should be implemented in all countries, but more as examples to show how others have used administrative data in statistical registers. The intention is to provide ideas that can be adapted to particular national circumstances rather than ready-made solutions.

1) Combining Multiple Sources

Figure 7.1 below is a simplified model of the sources used to maintain the statistical business register in the United Kingdom. It deliberately shows the statistical register at the centre, as the tool to combine and reconcile the data from the various sources. It also introduces the concept of satellite registers, which will be discussed in detail later in this chapter, and the idea that sources may already be a mixture of administrative and statistical data. In this case the geographic information system (GIS) already contains a mixture of administrative data (mainly from the postal service), with some statistical modelling, using population census data to create more statistically homogeneous areas.

Figure 7.1 – A Simplified Model of Statistical Business Register Sources in the UK



2) Using Centralised Administrative Registers

Centralised administrative registers are often created to improve efficiency within government, and in many cases they provide a single interface through which the subjects of the register can interact with different government agencies in a way that reduces duplication, and hence the burden of complying with administrative procedures. For example, where such a register exists, when a person or a business changes address, they only need to supply their new details once, and these details are then shared between all relevant agencies.

This sort of administrative register can be of immense benefit for statistical purposes, as it removes at least some of the burden of matching and reconciling data from different sources. To maximise the benefit, however, it is important for the statistical agency to have some say in the development and management of the administrative register, to ensure that it meets, as far as possible, statistical needs regarding units, classifications, definitions and procedures.

A good example of where this approach has worked in practice concerns the use of the (administrative) Australian Business Register (ABR)[5] by the Australian Bureau of Statistics. The ABR was developed by the Australian Tax Office to administer various businesses taxes, but is maintained in close cooperation with the Australian Bureau of Statistics, which provides input and expertise in specific areas such as economic activity classification.

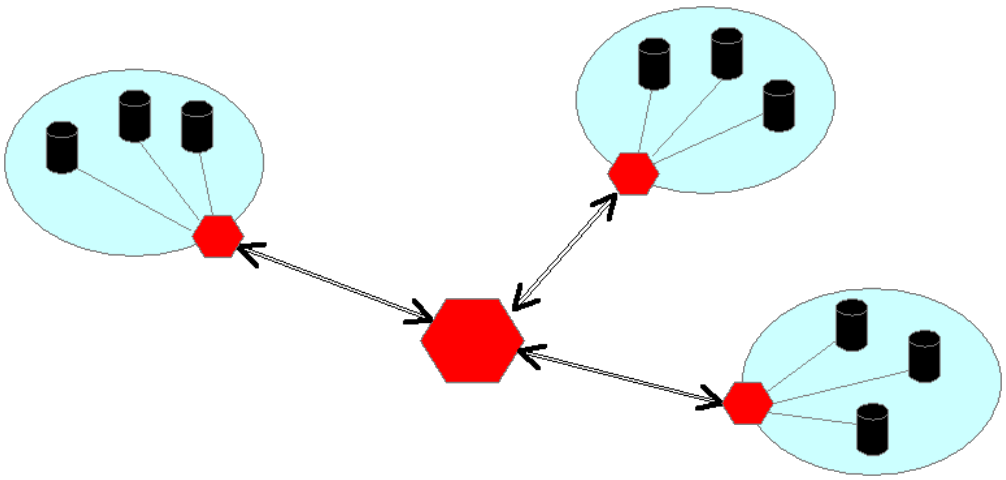
The result is that the ABR is a suitable basis for the statistical business register, for all but the largest and most complex businesses. In fact the statistical business register has a clear two-tier approach[6]. Most records are direct copies from the ABR, and are only maintained from that source, leaving statistical resources free to concentrate on maintaining the structures of the largest and most complex businesses.

3) Creating a Data-sharing Hub

A variation on the theme of a single centralised administrative register is the concept of a data-sharing hub. In this model, the central entity is not a fully fledged register, but is more of a tool for finding and matching data held by different agencies. It may contain some very basic identification data, but its main purpose is to provide a gateway through which data from different organisations can be shared within the government sector.

Figure 7.2 is taken from a study into the feasibility of such an approach in the UK[7]. This approach was not implemented, but the model remains a valid option for sharing administrative data. The blue circles represent different government bodies, each with a number of data holdings (the black cylinders). Each of these data holdings is linked to a portal which strictly controls what can pass through, and to whom. These portals are in turn linked to a central hub containing sufficient metadata to allow searching and matching of the linked data holdings. In this way, a user in one of the participating organisations can send a query via the central hub, and can receive data from all relevant data holdings in the other organisations to which that person has access rights.

Figure 7.2 – A Data Sharing Hub



4) Using Administrative Data via Satellite Registers

A rather different model for using administrative data in practice is to organise them into source-specific registers linked to a statistical register. If these source-specific registers meet certain criteria, they can be referred to as "satellite registers"[8]. Satellite registers can be defined as registers that are available to the national statistical system, contain information about units and variables of interest, and fulfil the following conditions:

- They are not an integral part of a statistical register, but are capable of being linked to it;
- They are more limited in scope than that statistical register, but within their scope they may have more extensive coverage of units and/or variables;
- They contain one or more variables that are not found in the statistical register. Such variables are generally capable of being used for stratification purposes;
- Databases in which results from surveys are normally recorded are not satellite registers

Satellite registers are therefore tools for incorporating administrative data that are only relevant for a sub-set of units in a statistical register. They may contain additional units, or variables, or both. They can be constructed using information from administrative sources, statistical surveys, or a combination of both. In some cases they may add, combine or otherwise transform variables, though in others they may be more or less identical to a particular source. To ensure that satellite registers are sufficiently coherent with statistical registers, it may be useful to consider additional criteria, e.g. common unit identifiers, common definitions and classifications. The greater the coherence, the more useful a satellite register is likely to be.

Figure 7.3 – The Relationship between a Satellite Register and a Statistical Register

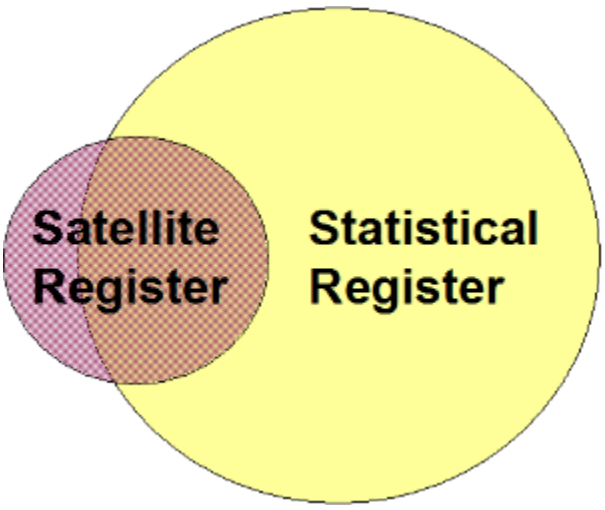


Figure 7.3 shows how a satellite register relates to a statistical register. This diagram can be interpreted both in terms of units covered and variables contained. In both cases there is a degree of overlap, but the satellite register also brings additional information, either additional units, or additional variables for a sub-set of existing units.

Most current examples of satellite registers relate to business data, where the scope of the satellite register can be determined by:

- Economic activity – the satellite register may contain businesses with specific activities, for example retail trade, hotels, road haulage etc.
- Size – The satellite register may contain units with a certain number of employees or turnover over a certain level, for example the subset of “large enterprises”

- Characteristics – The satellite register may contain units with a common characteristic, for example those that engage in foreign trade

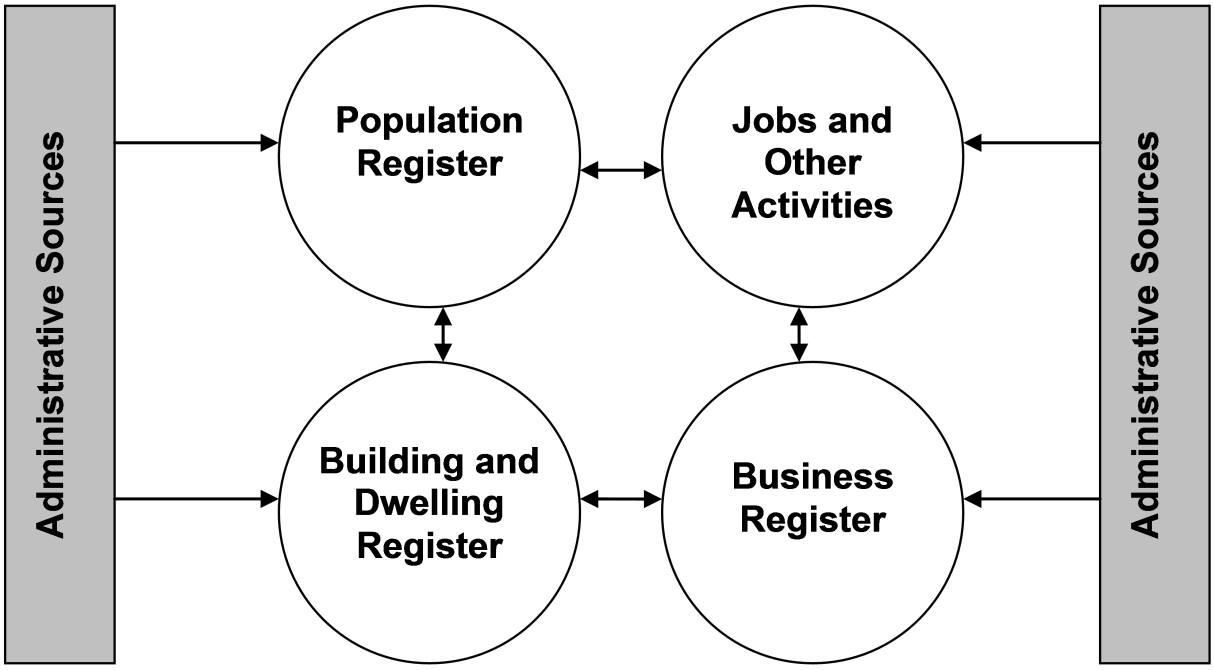
Examples of variables specific to the sub-set of units included in a satellite register could include “category” or “number of beds” for hotels, or “sales space” for retail businesses.

Satellite registers can add value to statistical registers by increasing the range of variables available for stratification and analysis purposes, and increase sampling efficiency by improving the quality of stratification variables. They may also increase the coverage of the target population, and in some cases can reduce the amount of information that needs to be collected via statistical surveys, thus reducing the burden on respondents.

5) Register-based Statistical Systems

Register-based statistical systems are discussed further in Chapter 9, but are mentioned here insofar as they offer a model for the use of administrative data in statistical registers. The main difference compared to the models described above is that several linked statistical registers are created using a wide range of administrative data. This model has been mainly developed in the Nordic countries, using either three or four core statistical registers. Figure 7.4 shows a simplified version of the model adopted in Sweden.

Figure 7.4 – Nordic Register-based Statistical Systems



The statistical population register is linked to a register of property or real estate, and to the statistical business register using a system of unique identifiers for people, properties and businesses. In Sweden, a fourth register has been introduced holding details about jobs or other activities. This register links people to their sources of income, including wages, pensions and state social security payments, and therefore shows the relationship between people and the labour market.

Annex to Chapter 7 – Exercise: Creating a Statistical Register of Entrepreneurs

Your government decides that it needs more data on entrepreneurs, and the factors that determine whether or not they are successful. Your office decides to produce a new data series to provide this information. You are asked to create a statistical register of entrepreneurs, based on administrative sources, to use as a sampling frame.

You have an annual budget of 16000 Euros. It costs 2000 Euros to process each data source that you use. In addition to this, there is the cost of buying the data, which varies from source to source.

The following administrative sources are available to you:

1. Tax office records of people that declare income from self-employment

- Contents: Person identification number, name, address, sex, amount of declared income, name of business, type of business (classified according to the International Standard Industrial Classification (ISIC) 2-digit level).
- Availability: The tax office will supply these data annually, if you pay a fee of 2500 Euros per year, to cover their costs of extracting and sending the data. They will send the data on CD-ROMs.
- Quality: The data are 95% accurate, except “type of business”, which is only 50% accurate. By the time you get the data, they will be

between 6 and 18 months out of date. Coverage is 100% of all people operating legal businesses. It is estimated that around 20% of businesses are operated illegally (i.e. by people who are not declaring their income).

2. Tax office records of businesses with employees

- Contents: Business identification number, name and address of business, number of employees, type of business (classified according to ISIC 4-digit level), year business first registered as an employer
- Availability: The tax office will supply these data if you pay an annual fee of 3000 Euros to cover their costs of extracting and sending the data
- Quality: The data are 90% accurate, and are typically between 2 and 3 months out of date. They will send the data monthly on CD-ROMs. They cover all businesses that are legally employing people. it is estimated that 50% of businesses have employees, and that 95% of these are operating legally.

3. Administrative population register

- Contents: Person identification number, name and address, age, sex, level of education, occupation, nationality, country of birth
- Availability: These data are already used by the statistical office, at an annual cost of 3000 Euros. If you use them, you would be expected to pay half of this cost. The data are available annually, and you can receive them as an electronic file from your colleagues in the population statistics division.
- Quality: The data are 95% accurate, but between 1 and 2 years out of date. They cover 99% of the legal population, but it is estimated that around 5% of the total population are illegal immigrants, so are not covered.

4. Telephone directory of businesses ("Yellow Pages")

- Contents: Name and address of business, telephone number, type of business (classified according to their own list of 300 categories)
- Availability: These data are sold commercially by a private sector company. They are available each month on CD-ROM. An annual subscription normally costs 7000 Euros, but the suppliers are willing to offer a discount of 15% to the statistical office.
- Quality: The data are claimed to be 99% accurate by the suppliers, who say that it is in the interests of businesses to make sure their information is correct. The data are typically between 1 and 2 months out of date. They cover around 85% of all businesses (legal and illegal).

5. List of people applying for business start-up grants

- Contents: Person identification number, name and address, business identification number, name and address, type of business (classified according to ISIC 2-digit level)
- Availability: 500 Euros for a spreadsheet sent by e-mail each March covering grant applications for the previous year.
- Quality: At least 95% accurate, though some addresses are out of date. Approximately 40% of people starting new businesses apply for a start-up grant, but these are typically the entrepreneurs that are most successful. This accounts for 6% of the total business population in any given year.

6. List of members of the "National Society of Entrepreneurs"

- Contents: Person name and address, business name, address and telephone number, date joined the Society
- Availability: 100 Euros for a paper directory published annually
- Quality: At least 90% accurate, though some addresses may be out of date. Membership fees are quite high, so only around 10% of entrepreneurs are members. These are mostly people with successful businesses that have been operating for at least 5 years.

Questions:

1. Given your limited budget (16000 Euros), which sources would you choose?
2. Why would you choose these sources?
3. How would you match the data from the different sources?
4. What type of survey would you recommend – personal interview, telephone interview or postal questionnaire?
5. Which variables would you use to stratify the sample for the survey?

Answers:

There are not really any right or wrong answers to this exercise, but the factors that should be considered include:

- Sources 1-3 are typical public-sector administrative sources, in that they have good coverage, but only of the legally registered units.
- Source 4 is a typical example of the type of private sector administrative data source that is increasingly being considered for statistical purposes in many countries. Note the possibility to negotiate on the price, there may be scope for further reductions, experience of negotiating commercial contracts would be useful!
- Sources 5 and 6 could be seen as typical satellite registers, in that they have limited coverage, but focus on a specific sub-population, which may have different characteristics to the population as a whole.
- Coverage, timeliness, accuracy and value-added should be considered as part of a cost-benefit analysis for each source.
- It would help to have more information about the user requirements for the resulting statistics, as this could influence the choice of sources. Experienced statisticians will recognize that requirements are often rather vague, at least initially, so further dialogue with users would be helpful. Issues for clarification could include:
    - Should the focus be on businesses that create jobs, or on the number of entrepreneurs?
    - What is the required balance between timeliness and accuracy?
    - Is there any user interest in attempts to estimate for entrepreneurs operating in the informal economy? If so, source 4 may be

needed, perhaps in combination with source 1.

Questions 4 and 5 are to some extent trick questions, as the initial response should be to see whether a survey is actually needed, or whether the required data can be produced directly from the statistical register created by combining the chosen sources.

---

[1] "Terminology on Statistical Metadata", UNECE / Conference of European Statisticians Statistical Standards and Studies, No. 53, Geneva, 2000, http://www.unece.org/stats/publications/53metadaterminology.pdf.
[2] See the Fundamental Principles of Official Statistics (Principle 6) http://www.unece.org/stats/archive/docs.fp.e.htm
[3] For example Austria - 'Bericht über die Einführung der Sektorklassifikation im Unternehmensregister der Statistik Austria' by Norbert Rainer, Karl Schwarz, Roland Schaumann and Thomas Karner. This paper contains an English summary, and is available on the Internet via the Eurostat restricted access 'BR-Net' site.
[4] See "System of National Accounts 2008", Chapter 4 - http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf
[5] See: http://www.abr.gov.au
[6] For more information see: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8165.0Explanatory%20Notes1Jun%202007%20to%20Jun%202009?OpenDocument
[7] For more information see: http://www.unece.org/stats/documents/ces/sem.46/5.e.pdf
[8] Sometimes also referred to as "associated registers".