Using Administrative and Secondary Sources for Official Statistics:

A Handbook of Principles and Practices

UNITED NATIONS

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

Using Administrative and Secondary Sources for Official Statistics

A Handbook of Principles and Practices

UNITED NATIONS
New York and Geneva, 2011

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontier or boundaries.

Acknowledgement

The UNECE would like to acknowledge the valuable contributions of more than two hundred participants of an international training course on the use of administrative sources for statistical purposes, and particularly the inputs from the presenters named below, all of which have helped considerably to enrich this handbook.

Mr Driss Afza, MEDSTAT projects
Ms Sue Fendall, Office for National Statistics, United Kingdom
Ms Riitta Harala, Statistics Finland
Mr John C Hughes, Office for National Statistics, United Kingdom
Mr Ben Humberstone, Office for National Statistics, United Kingdom
Mr Pekka Myrskylä, Statistics Finland
Ms Kaija Ruotsalainen, Statistics Finland

ECE/CES/13

Foreword

Statistical organisations around the world are coming under increasing pressure to improve the efficiency of the statistical production process, and particularly to make savings in costs and staff resources. At the same time, there are growing political demands to reduce the burden placed on the respondents to statistical surveys. This is particularly the case where respondents are businesses, as many governments see reducing bureaucracy as a key measure to support and promote business development.

Given these pressures, statisticians are increasingly being forced to consider alternatives to the traditional survey approach as a way of gathering data. Perhaps the most obvious answer is to see if usable data already exist elsewhere. Many non-statistical organisations collect data in various forms, and although these data are rarely direct substitutes for those collected via statistical surveys, they often offer possibilities, sometimes through the combination of multiple sources, to replace, fully or partially, direct statistical data collection.

The degree of use of administrative sources in the statistical production process varies considerably from country to country, from those that have developed fully functioning register-based statistical systems, to those that are just starting to consider this approach.

Although several subject specific texts exist, there have, until now, been no general, international methodological guidelines to help those in the early stages of using administrative data. This handbook aims to fill that gap. It builds on material developed over ten years in the context of an international training course on the use of administrative sources for statistical purposes. That course has now been delivered over ten times, to audiences of official statisticians from throughout Europe, Western and Central Asia, and North Africa.

Each time the course has been run, it has been improved and enhanced by sharing experiences with, and receiving feedback from participants. It has also benefited greatly from the input of various expert guest presenters from Statistics Finland and the British Office for National Statistics.

Mr Steven Vale, UNECE, Course leader

Contents

Notes

1) Note on References

This handbook includes many references to other papers, web sites and publications. To help those who want to follow-up these references, Internet addresses are given wherever possible. These were all checked at the time of writing, but that is no guarantee that they will still work at the time of reading. If the reader finds a broken link, please report it to support.stat@unece.org.

2) Note on Exercises

The exercises at the end of Chapters 6 and 7 are taken from the course on which this handbook was based. They are included here as practical examples to reinforce the theory presented in those chapters.

1. What are Administrative and Secondary Sources

1.1 Introduction

Before we start to consider the practicalities of using data from administrative and secondary sources, it is worth just taking some time to clearly define what these terms mean. Several definitions exist in the literature currently available, the most relevant of which are examined in this chapter. The chapter ends by proposing a relatively simple and broad definition, which is then used as the basis for the remainder of this handbook.

1.2 Traditional Definitions

Administrative sources have traditionally been defined as collections of data held by other parts of government, collected and used for the purposes of administering taxes, benefits or services. Perhaps the most comprehensive of the traditional definitions was set out by Gordon Brackstone of Statistics Canada in his 1987 paper "Statistical Issues of Administrative Data: Issues and Challenges"[1]. Brackstone identified four distinguishing features of administrative data:

1. The agent that supplies the data to the statistical agency and the unit to which the data relate are different (in contrast to most statistical surveys);
2. The data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;
3. Complete coverage of the target population is the aim;
4. Control of the methods by which the administrative data are collected and processed rests with the administrative agency.

This definition is broadly in line with that proposed by the Statistical Data and Metadata eXchange (SDMX) initiative[2]:

> *"A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations."*

During 1996-97 an internal Eurostat task force examined ways to better coordinate work relating to the use of administrative sources across different domains of statistics. This task force used a simple typology of data sources to consider how administrative sources should be defined. Firstly all data sources were divided into primary sources (data collected for statistical purposes) and secondary sources (all other data). A traditional or "narrow" definition of administrative sources comprises just public sector non-statistical sources, whereas a wider definition would also include private sector sources.

[1] Brackstone G J: "Statistical Issues of Administrative Data: Issues and Challenges", in "Statistical Uses of Administrative Data -An International Symposium", organised by Statistics Canada, 23-25 November 1987 (Proceedings published by Statistics Canada, Ottawa, December 1988).
[2] See: www.sdmx.org

The wider approach is consistent with the definition of administrative data adopted by the Conference of European Statisticians in the publication "Terminology on Statistical Metadata"[3]:

> *"Data collected by sources external to statistical offices."*

The narrow and wider definitions can be shown graphically as follows:

Figure 1.1 - Narrow definition

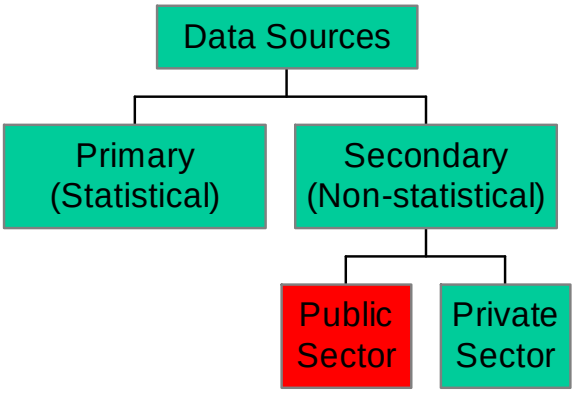
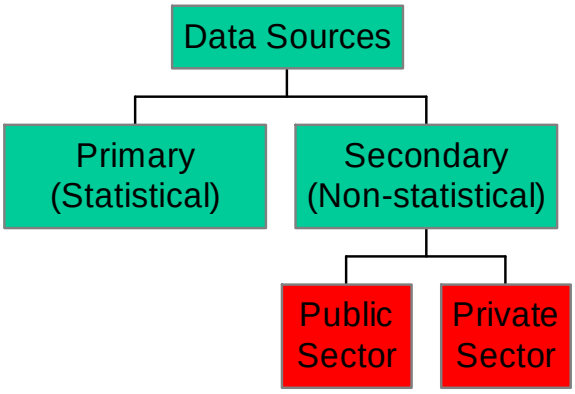


Figure 1.2 - Wider definition



Thus under the narrow definition, administrative sources are a sub-set of secondary sources, whilst under the wider definition these terms are synonyms.

There are a growing number of reasons for favouring the wider definition, including:

- ***Increasing privatisation of government functions:***

In several countries, regulatory functions that used to be carried out by government departments or agencies are being transferred to private or semi-private organisations. Typical examples are usually in the health, education or public utilities sectors, where former state monopolies are increasingly being replaced by private companies or non-profit institutions.

Registration functions, including the operation of administrative registers on behalf of government departments are also under consideration for privatisation in several countries. This means that the traditional distinctions between public and private sector functions are becoming increasingly blurred, and that the traditional or "narrow" definition of administrative sources is becoming too restrictive.

- ***Growth of private sector data and "value-added re-sellers":***

The amount of digital information in the world is growing exponentially, increasing by a factor of ten approximately every 5 years. Even if only a tiny fraction of this "data deluge" is of interest for official statistics, the volumes of data, and the range of topics they cover are still huge.

[3] See: www1.unece.org/stat/platform/download/attachments/9110092/Metadata+terminology+2000.pdf?version=1

At the same time, the commercial value of data is starting to become apparent, and the market for data is rapidly increasing within the private sector. This started with the development and sale of address lists for marketing purposes, it expanded to cover the provision of credit rating data and business intelligence information, and has now spread to cover virtually all types of data. As the size of this market has increased, so has the number of businesses seeking to profit from it. The private sector realises that data are a very valuable commodity.

A relatively recent development has been the emergence of private sector "value-added re-sellers" in the data market. These businesses take existing data from a variety of public and private sector sources, combine them, clean them, and sometimes validate them, and then re-sell them to other organisations. Examples include business data sellers such as Dun and Bradstreet, Bureau van Dijk and Hoppenstedt Bonnier.

This sort of data source can be of interest to official statistics providers, as it may be the case that these private sector data suppliers can actually process and supply data more cheaply than statistical organisations, often simply because they can spread the costs amongst a number of customers. The "Eurogroups" project to develop a European statistical register of enterprise groups uses such sources for exactly this reason.

An alternative to direct use of micro-data from such sources can be the use of aggregates for benchmarking purposes, comparing the coverage of target populations between private sources and official statistical registers. An exercise to compare the coverage of the UK statistical business register with that of leading private sector sources revealed statistical under-coverage of business activities in inner-city and holiday resort areas, illustrating the difficulties associated with covering marginal and seasonal activities in official statistics, as well as giving clear indications of the scale of this sort of under-coverage[4].

- ***User interest in new types of data***

Users of official statistics are constantly requesting new types of data. Pressures to reduce costs and burdens on respondents to statistical surveys make it difficult to launch new surveys to meet these demands, so statisticians increasingly need to look for alternative solutions. As the volume, content and coverage of private sector sources grows, so does their attractiveness as an alternative to statistical surveys.

1.3 Types of Administrative Sources

As discussed in the previous paragraphs, the potential range of administrative sources that could be used for statistical purposes is large and growing. The following list is not meant to be exhaustive; instead it aims to show range and types of potential data sources, as the final step towards arriving at an operational definition of administrative sources.

[4] The results of this exercise are shown in the form of a coverage map in the paper "The development of small area business statistics in United Kingdom", available at http://live.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf

- Tax data
  - Personal income tax
  - Value Added Tax (VAT)
  - Business / profits tax
  - Property taxes
  - Import / export duties
- Social security data
  - Contributions
  - Benefits
  - Pensions
- Health / education records
- Registration systems for persons / businesses / property / vehicles
- Identity cards / passports / driving licenses
- Electoral registers
- Register of farms
- Local council registers
- Building permits
- Licensing systems e.g. television, sale of restricted goods
- Published business accounts
- Internal accounting data held by businesses
- Private businesses with data holdings:
  - Credit agencies
  - Business analysts
  - Utility companies
  - Telephone directories
  - Retailers with store cards etc.

1.4 Summary

In conclusion, this chapter argues the case for a wide definition of administrative and secondary sources. It also highlights the need for imaginative assessments of the potential value of new types of data sources. For these reasons, the definition of administrative and secondary sources should not place any artificial restrictions on statisticians, and should be as wide as possible. As the terms “administrative sources” and “secondary sources” are therefore considered to be synonyms, this handbook will henceforth just use the term “administrative sources”, to cover both concepts.

The definition proposed is therefore:

> ***Administrative sources are data holdings containing information which is not primarily collected for statistical purposes.***

This definition is used as the basis for the contents of the rest of this handbook.

Box 1.1 – Looking to the Future: Store Cards – a Potential Data Source?

Store cards are a typical example of a new type of private sector data source. In return for benefits such as discounts and exclusive special offers, users of store cards give the stores a lot of data every time they use them. If you have a store card, the store knows or can derive the following data about you:

- Name, address, sex, age
- Family circumstances (e.g. if you regularly buy baby products, toys, pet food, or products such as meat in a certain quantity or size, it is easy to estimate the composition of your household)
- Indicators of work status and income (e.g. the time at which you shop can indicate whether or not you work, and the type of goods purchased can indicate disposable income)
- Other household indicators, such as car ownership (purchases of petrol and car care products), religion (purchase of goods linked to a particular religion, e.g. halal or kosher meat), etc.

This may seem a rather extreme example of a potential source, and one that is unlikely to be considered for the purposes of official statistics in the near future. However, several countries have considered the use of till roll data from major retailers as a source of data on retail sales and prices, and Statistics New Zealand has produced an experimental data series using electronic card transaction data[5].

The use of store card data could be seen as the next logical step, particularly if coverage can be improved by linking data from different store card schemes, as well as data from other commercial sources. If this sort of administrative data source is ignored by official statisticians, how long will it be before private sector businesses with access to these data, start to offer plausible, and more cost effective alternatives to key official statistical outputs such as population census data?

[5] http://www.stats.govt.nz/browse_for_stats/Corporate/Corporate/nzae-2007/~/media/Statistics/Publications/NZAE/The%20development%20of%20electronic%20card%20transaction%20statistics/development-of-ect-statistics.ashx

2. The Advantages of Using Administrative Sources

2.1 Introduction

The previous chapter defined the nature and scope of administrative sources, but did not really consider why these sources are of interest to statisticians. This chapter considers the many potential benefits of using administrative sources in official statistics, either to complement or replace statistical sources. Of course, it is not all good news, along with the benefits there are also usually a range of problems to be overcome. These problems, and how they might be solved, are dealt with in Chapter 4.

2.2 Cost

Statistical surveys are an expensive way of collecting data. Questionnaires have to be developed, samples have to be designed (which may even require the creation of a specific sampling frame), respondents have to be contacted, and possibly re-contacted to encourage them to reply, responses have to be processed and verified, and results have to be calculated. Although computers can take much of the processing burden, the whole approach is still rather labour intensive, particularly the response chasing stage, which can probably never be fully automated.

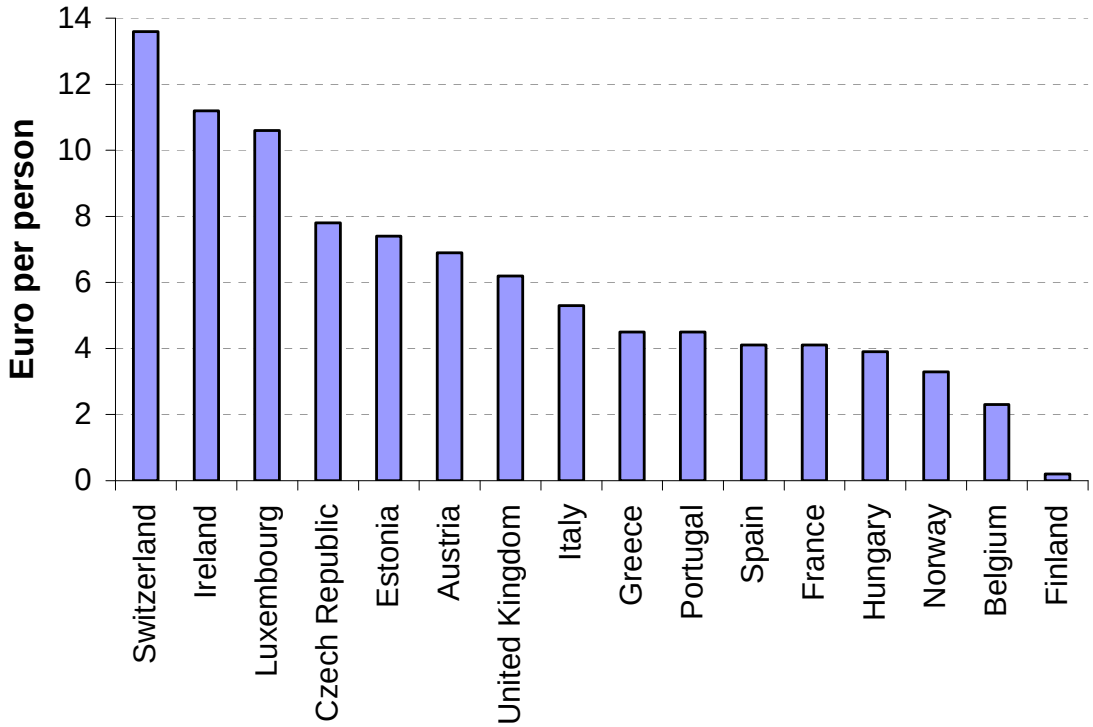
Traditional censuses are even worse because they are conducted on a much larger scale. National statistical organisations that still conduct traditional censuses of people, businesses, farms etc., often require special funding for such exercises, as they are too costly to be covered within their regular budget. This makes traditional censuses highly visible to politicians and therefore vulnerable to changes in political priorities.

Although the set-up costs of using administrative sources to produce statistical outputs can easily be as high as the set-up costs for a statistical survey, the running costs are usually significantly lower. Table 2.1 and Figure 2.1 below show the costs of conducting population censuses in 2000-2001, in European Union countries. The huge differences in the cost per head of population between Finland, where the census was totally based on administrative sources, and other countries such as the United Kingdom and Austria, where traditional paper questionnaires were used, is perhaps the strongest argument available for greater use of administrative data.

Table 2.1 - Population census costs in selected European Union countries

Country	Total cost (millions of Euro)	Cost per person (Euro)
Belgium	24	2.3
Greece	50	4.5
Spain	167	4.1
France	248	4.1
Ireland	44	11.2
Italy	298	5.3
Luxembourg	5	10.6
Austria	56	6.9
Portugal	46	4.5
Finland	0.8	0.2
United Kingdom	367	6.2
Norway	15	3.3
Switzerland	99	13.6
Czech Republic	80	7.8
Estonia	10	7.4
Hungary	40	3.9

Figure 2.1 - Comparative population census costs per person



Source: Table 22 of the Eurostat publication: "Documentation of the 2000 round of population and Housing censuses in the EU, EFTA and Candidate Countries"[6]

[6] Available from Eurostat at http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-CC-04-002/EN/KS-CC-04-002-EN.PDF

Access to administrative sources is often free of charge, particularly if the data originate from the public sector. Even if there is a charge, for example to cover data extraction or transmission costs from a public source, or to buy data from a private source, it is often still cheaper to use administrative data than to collect the same information via a survey.

Where statistical surveys are still used, an efficient and accurate sampling frame is needed. The statistical registers used to produce those sampling frames are often so large and complex that it is very difficult and expensive to satisfactorily populate and maintain them using survey or census data. Therefore even if administrative data do not replace statistical surveys, they can still be used to populate and maintain statistical registers, and thus help to reduce overall costs.

2.3 Response Burden

Using data from administrative sources helps to reduce the response burden on data suppliers. This is a strong political consideration in many countries, particularly if the respondents are businesses. Policies to encourage business development and growth often include reducing regulatory burdens. In these circumstances, statistical surveys are often seen as an easy target for cuts.

From their side, businesses usually understand the reasons for supplying data for registration and taxation purposes, even if they do not like doing so. They often, however, see statistical data requests as an extra, less necessary, burden. If they have already provided details to other government departments, they may become annoyed at receiving similar requests from the national statistical organisation. Thus, if policy makers and respondents are united in calling for reductions in the statistical response burden, it is extremely difficult for national statistical organisations to resist this pressure, and the re-use of data collected by others is the logical solution.

2.4 Frequency

Related to the reductions in cost and response burden, a further advantage of the use of administrative sources is that they may in some cases allow statistics to be produced more frequently, with no extra response burden, and little extra cost. This is the case in Finland, where it is possible to produce population census data from administrative sources on an annual basis, whereas countries using more traditional methods can only afford to produce these data every five or ten years.

The main constraint to the frequency of statistics produced from administrative data is usually the frequency with which the administrative source is updated. Thus it would be difficult to produce monthly statistics from administrative data updated once per year, unless those data were updated on a rolling basis with no seasonal bias (or at least sufficient information to remove any seasonal biases).

Administrative sources that are not based on any particular time period, such as those that record events (e.g. birth, death, granting of planning permission), however, offer considerable flexibility. This is because, as long as the date of the

event is recorded accurately, they allow statistics to be produced for any given period or periodicity down to daily.

2.5 Coverage

Administrative sources often give complete, or almost complete, coverage of their target population, whereas sample surveys can often only directly cover a relatively small proportion directly. The use of administrative sources therefore eliminates survey errors, removes (or significantly reduces) non-response, and provides more accurate and detailed estimates for various sub-populations, e.g. respondents in small geographic areas, or with other specific characteristics.

2.6 Timeliness

The use of administrative sources may increase the timeliness of statistical outputs by allowing access to more up to date information concerning certain variables. This is because statistical surveys generally take time to plan, to design and pilot forms, to analyse the population and optimise the sample etc.. This is particularly the case for annual or ad-hoc data collections. Therefore access to a suitable administrative source can be a more efficient solution. It should be noted, however, that there are also likely to be cases where the use of administrative sources leads to a reduction in timeliness, particularly regarding short-term indicators.

One area where administrative sources can have a particularly positive impact on timeliness is in the management of statistical registers and survey frames. Administrative information on changes to the target population (e.g. births and deaths of people or businesses) is often much more up to date than survey information could ever be, simply because of the coverage benefits mentioned above.

2.7 Public Image

Public opinion relating to the sharing of data, particularly between different government departments, varies considerably from country to country. Where public opinion generally accepts, or is in favour of data sharing, the increased use of existing data sources can help to enhance the prestige of a national statistical institution by making it more efficient and cost-effective.

Although there is often a general unease amongst the public about data sharing, there are also contradictory pressures to improve the efficiency of government, particularly if this results in lower taxes or more funding for voter-popular areas such as health or education. Political slogans such as “joined-up government” are often appealing to the public, and can help to counter fears of loss of privacy. Thus the extent to which improvements to public image can be seen as an advantage of using administrative sources depends heavily on how that use is presented to and perceived by the public.

3. Frameworks for Access to Administrative Sources

3.1 Introduction

The access to data from administrative sources is one of the key barriers to the wider use of such data for statistical purposes. This chapter describes the various frameworks needed to facilitate access to administrative sources, drawing on examples and experiences from several countries. These frameworks typically have several dimensions; legal, policy, organisational and technical, each of which is considered below. It is necessary to reach agreement in all of these areas before the benefits of the use of administrative data can be realised.

3.2 Legal Frameworks

Legal frameworks are normally constructed at the national level, and are specific to national sources and circumstances. In some cases, however, there may also be relevant legislation at either the sub-national (e.g. state) level, or the international level. An example of the latter is the statistical legislation of the European Union, which is binding on Member States. In such cases, it is possible that there are two or more alternative legal gateways to administrative data.

Most national statistical organisations have legal texts defining their roles and responsibilities, typically in the form of a statistics act. In many countries, these legal texts include specific provisions for the access to administrative data. Examples include the statistics acts of Ireland[7] and Norway[8].

Box 3.1 – Extracts from the Irish Statistics Act of 1993

Section 30. (1) For the purpose of assisting the [statistical] Office in the exercise of its functions under this Act, the Director General may by delivery of a notice request any public authority to –

> (a) allow officers of statistics at all reasonable times to have access to, inspect and take copies of or extracts from any records in its charge, and
> (b) provide the Office, if any such officer so requires, with copies or extracts from any such record,

and the public authority shall, subject to subsection (2) of this section, comply with any such request free of charge.
.....

Section 31. (1) The Director General may request any public authority to consult and co-operate with him for the purpose of assessing the potential of the records of the authority as a source of statistical information and, where appropriate and

[7] www.irishstatutebook.ie/1993/en/act/pub/0021/index.html (See sections 30 and 31)
[8] www.ssb.no/english/about_ssb/statlaw/statlov_en.html (See chapter 3-2)

practicable, developing its recording methods and systems for statistical purposes, and the public authority shall comply with any such request, in so far as resources permit.

(2) If any public authority proposes to introduce, revise or extend any system for the storage and retrieval of information or to make a statistical survey it shall consult with the Director General and accept any recommendations that he may reasonably make in relation to the proposal.

Some national legal frameworks give more powers than others for access to administrative data for statistical purposes. This is because national historical, political and cultural factors have a strong influence on these frameworks. Cultural factors can be particularly important, as some cultures are much more favourable than others to the idea of data sharing between government departments and agencies. As a result of these national differences, legal frameworks are not particularly harmonised or even consistent between countries.

To address this issue of consistency, the European Union has included provisions on access to administrative data in Regulation 223/2009 on European statistics, commonly known as the "Statistical Law"[9]. This Regulation gives the national statistical organisations of Member States the right of access to the administrative data needed to meet their obligations under European statistical legislation, but states that such access is still subject to national limits and conditions.

Individual European Union regulations in specific areas of statistics go further, and remove this dependency on national limits and conditions. An example of this is the business registers regulation, which gives unconstrained access to any administrative sources, where data from these sources are necessary to meet the requirements of the regulation[10].

As well as giving access to data from administrative sources, legal frameworks also set out the limits to such access, and to the uses of administrative data. Often there are restrictions that data can only be used for specific statistical purposes, and that the confidentiality of individual records should be maintained.

For example, there may be specific restrictions on the use of data for unincorporated businesses, particularly sole-proprietorships, where business data could be considered to be personal data relating to the owner of the business. In such cases, business profit can be seen as equal to personal income. Many countries have data protection legislation covering information about individual citizens, therefore it is important to make clear distinctions between what constitutes business and personal data in such cases.

[9] Article 24 of Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:EN:PDF.

[10] Article 4 of Regulation (EC) No 177/2008 of the European Parliament and of the Council of 20 February 2008 establishing a common framework for business registers for statistical purposes: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:061:0006:0016:EN:PDF

The legislative process can take time, and statistics may often be seen as a relatively low priority by legislators, so a sustained period of lobbying and highlighting the benefits of using administrative data may be necessary. Given all the efforts that are usually needed to introduce or revise statistical legislation, it is therefore necessary to make the most of the opportunity. In particular, it is essential to avoid the mistake of proposing legislation that just meets current requirements. It may be ten or more years until the next opportunity to revise legislation, so it is necessary to have a long-term strategy for the use of administrative data, and to ensure that the legislative proposals meet all envisaged requirements for the foreseeable future. In this way, legislation can be seen as a short-term barrier, but a long-term opportunity.

Even whilst legislation remains a barrier, it does not necessarily prevent any use of administrative data. In one example, whilst waiting for a suitable legal framework for access to Corporation Tax data, a member of staff from the United Kingdom Office for National Statistics was seconded to the tax agency to explore the feasibility of using these data for statistical purposes. This person had access to the micro-data whilst seconded to, and physically working in the premises of the tax agency, but could only take non-disclosive, aggregate analyses back to the statistical office. This approach meant that various data issues could be addressed, including a proper assessment of the real value of the tax data, whilst simultaneously exploring the possible legal routes to gaining access.

It should also be noted that legislative restrictions often concern the use of micro-data, i.e. information on individual people or businesses. Although statisticians are habitually used to working with data at this level to produce aggregate results, it may sometimes be feasible to work with non-disclosive, low-level aggregates instead. In some cases, this could be done by simply re-defining the statistical unit from the individual to a small group of individuals sharing certain characteristics, perhaps with a weight equal to the number of members of that group.

3.3 Policy Frameworks

Many countries have general policies on data sharing within government, which will influence the right of access to administrative data for statistical purposes. However, it is often easier to change policies than to change laws, and policy tends to evolve over time. It is therefore important that national statistical organisations participate fully in policy development, and take an active part in any discussions within government that might lead to policy changes. In this way, any changes should be formulated in a way that gives the maximum possible benefit to the statistical system.

Policy frameworks also encompass voluntary codes of practice, the most important of which, for statistical purposes, is the United Nations “Fundamental Principles of Official Statistics”[11]. Principle 5 concerns cost-effectiveness, and suggests the use of data from administrative sources in this context:

[11] http://unstats.un.org/unsd/methods/statorg/FP-English.htm

"Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents."

An explanatory note to principle 5 also stresses cost-effectiveness, and goes on to say that:

"Statistical offices must be cost-effective, making the best choice of concepts, sources and methods by balancing quality, timeliness, costs and the reporting load of respondents ..... The overall cost-effectiveness of an agency is influenced by organizational planning and operation, the sound application of statistical methodology, exploitation of information and communication technology and also access to administrative records."

The code of practice for the European Statistical System[12] contains similar provisions, but the use of data from administrative sources is encouraged in slightly different contexts. Principle 2 concerns the mandate for data collection, and states that:

"Statistical authorities must have a clear legal mandate to collect information for European statistical purposes. Administrations, enterprises and households, and the public at large may be compelled by law to allow access to or deliver data for European statistical purposes at the request of statistical authorities."

Principle 9 is concerned with ensuring that the burden on respondents to statistical surveys is not excessive. It states that:

"The reporting burden should be proportionate to the needs of the users and should not be excessive for respondents. The statistical authority monitors the response burden and sets targets for its reduction over time."

One of the proposed indicators to measure the application of this principle is:

"Administrative sources are used whenever possible to avoid duplicating requests for information."

Codes of practice may also exist at the national level, and are often valuable as a way of reassuring the public that data will only be used for specific and reasonable purposes. To have any real value, it is important that these codes of practice are made available to the general public, typically via the internet site of the national statistical organisation.

3.4 Organisational Frameworks

Once the legal and policy frameworks are in place to permit the use of administrative data, it is necessary to consider the organisational arrangements to facilitate data

[12] http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice

flows. Typically this takes the form of a written agreement. This may be a contract, particularly if a private sector organisation is involved, but, if the agreement is between government departments or agencies, it is more likely to be a “service level agreement”, “protocol” or “concordat”. The difference is that contracts tend to be legally binding, whereas other forms of agreement are not.

There are certain key features that should be present in any such agreement. These are as follows:

- The legal basis: A reference to the legislation permitting the access to the administrative source for statistical purposes, and to any legislation that imposes restrictions on this access.
- Names of persons transferring / receiving data: The names, contact details and job titles of the key people involved in the transfer of data in both the administrative and statistical organisations should be recorded. In some cases, this can include all persons within the statistical organisation who are entitled to use or view the data.
- Detailed description of data covered: This will include information identifying the data set and the variables contained within it.
- Frequency of data supply: This will specify when and how frequently the administrative organisation will supply data.
- Quality standards: These set the parameters for the quality of the data supplied. Examples might include requirements for addresses to meet certain standards, or for a maximum proportion of missing or erroneous variables, to ensure that the data received are fit for purpose. The priorities assigned to different variables, and hence the effort put into quality assurance, will often differ between administrative and statistical organisations, therefore agreeing common standards can be difficult.
- Confidentiality rules: It is important to set out what the data can be used for, what rules and procedures will be in place to prevent disclosure, and in what circumstances the data can be passed on to customers of the statistical organisation.
- Technical standards: These are covered in more detail in the technical frameworks section below.
- Provision of metadata: It is important that data flows are accompanied by the relevant metadata, which can include dates, descriptions for any codes used, information on the units used, etc.
- Provisions for payment for supply of data: Data transfers between government departments or agencies are generally free of charge, though in some cases, the statistical organisation may be required to contribute towards the costs of extracting and transferring the data. Data from private sector organisations are often charged for at market rates, though it may be possible to negotiate discounts, particularly if there are several users of a private sector data source within government. In some cases, it might be possible to offer statistical analyses or expertise as a form of payment for data received.
- Period of agreement: Agreements may be for a fixed period, but if so, they should include provisions to renew or extend them as necessary. An alternative approach is to have an agreement that is considered valid until one of the parties wants to make a change.

- Contingencies for changes in circumstances: It is important for the statistical organisation to have advance warning of changes affecting the administrative source. The agreement should specify that any proposed changes are communicated to the statistical organisation as soon as possible, to allow the impact of the changes on statistical outputs to be minimised.
- Procedure for resolving disputes: The agreement should specify the method to resolve any disputes between the statistical and administrative organisations, which will normally be to escalate issues to senior managers, or possibly even to the relevant ministers.

3.5 Technical Frameworks

Technical frameworks refer to the mechanisms by which data are transferred, as well as any relevant data or metadata standards. Data transfer mechanisms can take any form from paper records sent by post to real-time updates via a secure electronic link. The mechanism used has to take into account the technical possibilities open to both the sending and the receiving organisation, so is often a compromise reflecting a sub-optimal solution for at least one of these organisations.

There are a number of international standards for data and metadata transmission, including XML, SDMX and DDI, to name but a few. Some countries also have national versions, particularly for data transfers within government. It is therefore important to agree which standards are to be used.

3.6 Summary

It is essential to have a legal framework in place to permit the use of administrative data for statistical purposes. The other frameworks described above are not essential, but are very useful for assuring a smooth flow of data, and minimising any problems or misunderstandings between the data supplier and the statistical organisation. For this reason, it is helpful if they are reflected in written documents, agreed by all parties.

Comparing practices between countries can be useful for benchmarking purposes, but it should be remembered that specific national situations and issues often require specific solutions. International standards can help in terms of providing guidance (and compliance with them may be seen as a political goal) so they should be quoted wherever possible in discussions with administrative departments.

Box 3.2 – Case Study: Frameworks in the United Kingdom

- ***Legal frameworks***

The Statistics and Registration Services Act 2007 provides the framework for access to administrative data, but does not give a blanket right of access as in many other countries. The conditions of access and use of administrative data are often governed by source-specific legislation such as the Value-added Tax Act of 1994. Access to new administrative sources is subject to parliamentary approval. As a member of the European Union, the United Kingdom is also subject to the provisions of the European legislation relating to the use of administrative sources.

- ***Policy Frameworks***

In addition to applying the United Nations Fundamental Principles of Official Statistics and the European Statistical System Code of Practice, there is a national code of practice[13] for members of the Government Statistical Service. The key provisions in relation to the use of administrative data for statistical purposes are:

- *"5(f) The same confidentiality standards will apply to data derived from administrative sources as apply to those collected specifically for statistical purposes."*
- *"7(c) The value of administrative data in producing National Statistics will be recognised, and statistical purposes should be promoted in the design of administrative systems."*
- *"7(d) Statistical systems will be designed in ways that maximise the potential to add value through data integration."*

The code of practice is supported by various protocols, including a Protocol on Managing Respondent Load[14], which contains the following statements:

- *"2. New statistical surveys will not duplicate existing sources... Producers of National Statistics will consider using existing survey data, administrative data and other non-survey sources before introducing a new survey... A survey will be conducted only where there is no suitable alternative data source."*
- *"4. The value of administrative data in producing National Statistics will be recognised, and statistical purposes should be promoted in the design of administrative systems. National Statistics will, where appropriate, be derived from information supplied for the administration of government business and public services. This will be achieved, wherever possible, by direct extraction of relevant data from the systems supporting the administration. Producers of National Statistics will seek to influence those responsible for the design of administrative systems so that these systems can also capture data for statistical purposes in an economical way."*

[13] http://www.statistics.gov.uk/about/national_statistics/cop/default.asp
[14] http://www.statistics.gov.uk/about/national_statistics/cop/downloads/respondentload.pdf

- ***Organisational Frameworks***

The organisational frameworks for the transfer of data between government departments and agencies tend to be incorporated in “service level agreements”. These are signed at a senior level, but are not legally binding. They contain general provisions in the main part of the agreement, and have details of specific data requirements and specifications in annexes. Usually there is no payment, but in some cases, statistical analyses or tools are provided in return.

Government departments and agencies that supply administrative data for the statistical business register are represented on the management committee for that register, which also includes users. This helps them to better understand how their data are used, and the implications of data quality.

The company registration agency (Companies House) operates on more of a commercial basis, so the framework for the transfer of data from that agency takes the form of a contract, with a payment. Data on business ownership and control links are also purchased from a private sector business data supplier.

- ***Technical Frameworks***

Most data transfers are via text files, with either fixed-length fields or standard delimiters. One area where some standardisation has been possible is in the format of business addresses. This has been facilitated by the availability of address referencing software tools based on Post Office standards.

Most data transfers are currently via discs sent by post, or, for smaller data-sets, via secure e-mail links. However, for Value Added Tax data used in the statistical business register, a system of daily updates has been set up using transaction files sent via the government secure intranet.

Metadata are usually transmitted as reference tables, either accompanying the data, or separately, on a less frequent basis. Metadata defining codes are stored as look-up tables, whereas more general metadata are recorded in a database of standards and guidance.

4. Common problems and solutions

4.1 Introduction

Although Chapter 2 outlined many good reasons for using administrative sources, there are also a number of problems associated with their use. Some of these problems are specific to a particular source or use, but many of them will be more generic in nature. This chapter outlines some of the more common problems and proposes methods to solve them, or at least to minimise their impact. The specific problems of getting access to administrative sources in the first place, and of data linkage, are treated separately in Chapters 3 and 6 respectively.

4.2 Public Opinion

Chapter 2 considered how public opinion might favour the sharing of data in some countries. In other countries, however, there may be public unease at the thought of data being shared around government. It is very difficult to reduce such concerns, but possible approaches could include the publication of clear limits and rules regarding the use of data, ensuring that people and businesses understand that sensitive data used or collected for statistical purposes will not be fed back to other parts of government (particularly tax and benefits agencies).

This is in line with the United Nations "Fundamental Principles of Official Statistics", where Principle 5 ("*Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents*") encourages the use of administrative data. Taken together with Principle 6 ("*Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes*"), this establishes the principle of the one-way flow of data.

Other ways to help overcome hostile public opinion include the publication of analyses of the costs and benefits, both to government and to respondents, of the use of different sources. It may also be possible to claim that micro-data are more secure when administrative sources are used. No questionnaires are sent by post, data are not held on paper or electronically by interviewers, and fewer clerical staff are needed for the statistical production process, thus fewer people have access to sensitive data.

4.3 Public Profile

Direct contact with the public via surveys helps to raise the profile of the statistical organisation. The use of administrative data can reduce that contact and hence also reduce public awareness of the work of the statistical organisation. If this becomes an issue, the most obvious solution is to improve the 'marketing' of the statistical

organisation and its data outputs. This may require a small proportion of the savings from using administrative sources to be transferred to the marketing budget.

Perhaps the most effective way of promoting the activities and outputs of a national statistical organisation, particularly in the medium to long term is to ensure greater involvement with education institutions, business groups, and other target customers. User groups are also particularly important in this respect, and should be actively encouraged.

4.4 Management of Change

Public sector administrative sources are generally set up for the purposes of collecting taxes or monitoring government policies. This means that they are susceptible to political changes. If a policy changes, administrative sources may be affected in terms of coverage, definitions, thresholds etc., or possibly even abolished completely. Changes to the computer systems used to store and process administrative data may also have an impact on the supply of data for statistical purposes. Even private sector sources are not immune from these sort of changes, though in this case, change is more likely to be driven by changing market factors

Such changes may happen suddenly, with little warning, particularly high-risk times tend to be immediately after a change of government, a change of minister, or a change in legislation. An example was reported some years ago from Slovenia, where the supply of administrative data on employment was halted for a while following a change of minister, leaving the statistical office with serious problems for the production of employment statistics. Procedures, backed up by legislation, have since been implemented to minimise the likelihood and impact of this sort of change.

Reliance on a particular source will always, therefore, carry a certain degree of risk. These risks can be managed to some extent by legal or contractual provisions. The best way in practice to avoid such problems tends to be through regular contact with those responsible for the administrative source, to ensure they are aware of the statistical requirements, and to try to influence and get early warning of any possible changes. Where there is a strong reliance on a particular source, it is also worth preparing contingency plans setting out what could be done if that source became unavailable It is clearly better to be proactive beforehand than have to react after the event!

4.5 Units

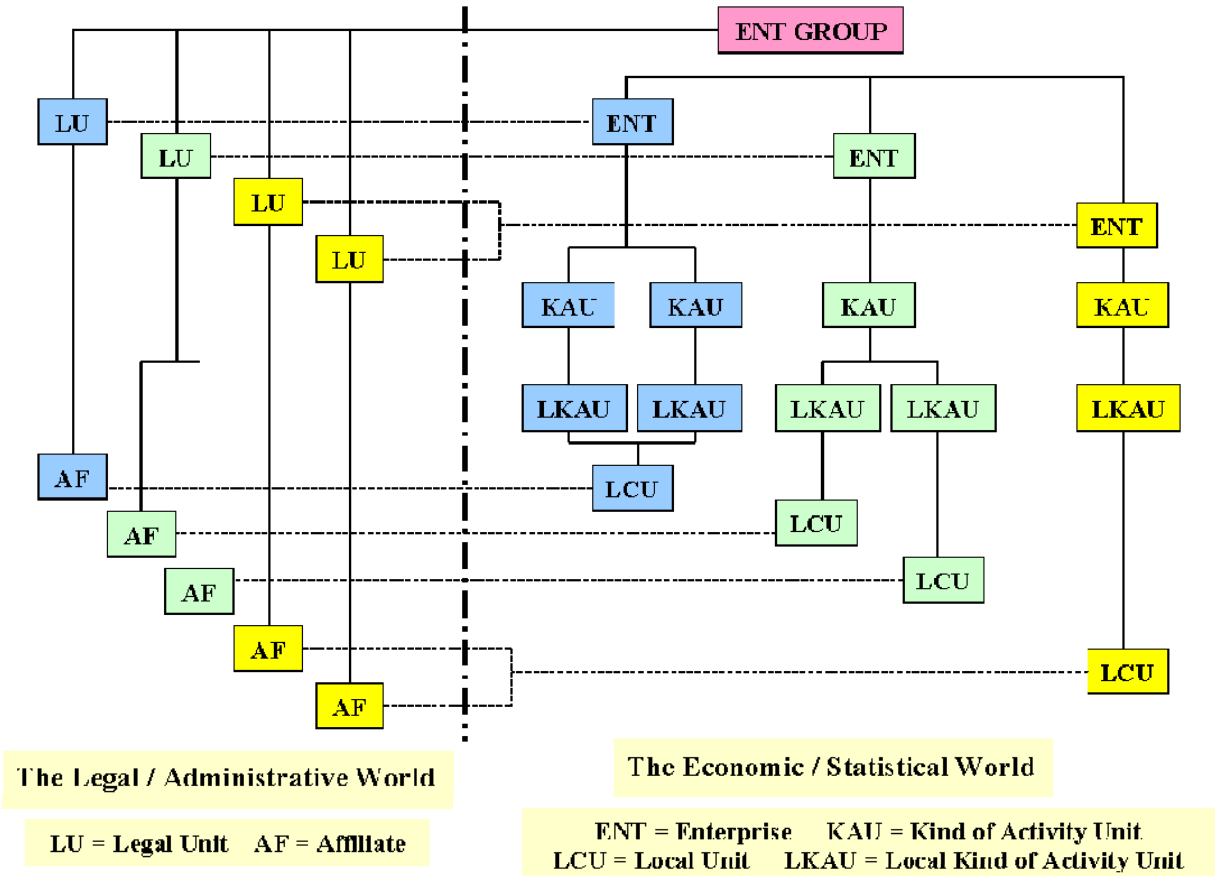
One major problem often encountered when using administrative sources is that the units used in those sources do not correspond directly to the definition of the required statistical units. The process of converting from administrative units (legal units, tax units, claimants etc.) to statistical units (enterprises, people, households etc.) can be quite difficult conceptually, and often involves some form of modelling.

In business statistics, this process is known as profiling, and typically is a function of statistical business registers. Eurostat has published guidelines for this process in

Chapter 19 of their Business Register Recommendations Manual[15], where they define profiling as "*a method to analyse the legal, operational and accounting structure of an enterprise group at national and world level, in order to establish the statistical units within that group, their links, and the most efficient structures for the collection of statistical data.*"

Figure 4.1 shows how the structure of a set of linked business units can look very different from the legal / administrative point of view, compared to the statistical point of view. Profiling, as defined above, can be seen as the process of creating the statistical structure and mapping it to the legal / administrative structure.

Figure 4.1 – Different views of a group of business units



Although profiling gives a better understanding of complex unit structures, it is expensive and time consuming, and needs trained staff. It is therefore totally impractical to attempt detailed clerical profiling for all business units in an economy, it is necessary to focus on those cases that give the most benefit. Profiling can be seen as a trade-off or compromise between three factors:

- Quantity of business structures profiled;
- Quality or depth of the profiling activity and;
- Available resources (determined both by cost and suitability of staff).

[15] Available from Eurostat at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-BG-03-001/EN/KS-BG-03-001-EN.PDF

Box 4.1 gives four examples of business structures that were profiled separately in three different countries (Denmark, The Netherlands and the United Kingdom) as part of a study into the consistency of application of the statistical definition of the enterprise within the European Union[16]. This shows clearly that profiling is to some extent an art, and there is not always a "right" answer, however this particular exercise resulted in considerable methodological work to harmonise the rules for profiling, which is partially documented in Chapter 19 of the Eurostat Business Register Recommendations Manual cited above.

Although clerical profiling is not practical for all units in a large population, some form of automated, rules-based profiling might be. Standard rules based on attributes or the nature of links between units can help to overcome differences between administrative and statistical units in many areas of statistics. For example, statistical households can be derived based on relationships between the individuals living in a building. This approach is used successfully within the register-based population census methodology applied in Nordic countries.

An alternative to profiling that may be feasible in some cases is to consider correcting for differences in the definitions of units by making statistical "adjustments". A crude example of this approach could be where the statistical unit is persons and the administrative unit is jobs. Assuming that it is known from a survey that working people have, on average, 1.15 jobs, this adjustment factor can be used to estimate persons in employment from the number of jobs.

Box 4.1 – An Exercise in Profiling: How Many Enterprises?

The following examples are taken from the study "The Impact of Diverging Interpretations of the Enterprise Concept", prepared for Eurostat by Statistics Netherlands with input from Denmark and UK. Each example is followed by the answer given by each of the three participating countries, along with a summary of their reasoning. The examples are based on the following definition of an enterprise:

> "*the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making .... An enterprise may be a sole legal unit.*"
>
> Source: EU Regulation 696/93 on statistical units

Example 1 - Two legal units in an enterprise group have different 4 digit NACE codes; both are selling mainly to third parties outside the group. They share buildings, management, purchases and employees.

Answers

- Netherlands and United Kingdom: Combine into one enterprise, given the intensity of shared production factors.

[16] "The Impact of Diverging Interpretations of the Enterprise Concept" - a study prepared for Eurostat by Statistics Netherlands with input from Denmark and UK.

- Denmark: Two separate enterprises, as both sell more than 50% outside the group.

Example 2 - Four Legal Units: A and B have different activities, no combined purchases, but share buildings. C and D share buildings, employees, and purchases. All four present themselves as one firm.

Answers
- Netherlands and Denmark: A and B are separate enterprises, combine C and D into one enterprise, because A and B operate on market terms, whilst C and D share production factors.
- United Kingdom: All four in one enterprise because they present themselves as one firm

Example 3 - Three legal units: All produce mainly for external customers, they share management and purchases, and represent themselves as one firm. A and B share a building. B and C have the same activity, share employees and capital goods and can not supply separate data.

Answers
- Netherlands: Combine into one enterprise because all share management and purchases, and represent themselves as one firm.
- United Kingdom and Denmark: Combine B and C into one enterprise, because they are horizontally integrated, and data are only available for these two together. A is a separate enterprise

Example 4 - Twelve legal units form an enterprise group. Only one is active, the others have no employees.

Answers
- Netherlands: One enterprise which only consists of the active unit, because units which are not active are not part of an enterprise.
- United Kingdom: One enterprise which consists of all units, because there is no point having separate enterprises for non-active units.
- Denmark: Each unit is a separate enterprise, because there are no strong ties between the units

4.6 Definitions of Variables

As well as differences in the definitions of units, there are also likely to be differences in the definitions of variables between administrative and statistical systems. The data in administrative sources have generally been collected for a specific administrative purpose, and the needs and priorities relating to that purpose are likely to be different to those of the statistical system. For example, turnover for value added tax (VAT) purposes may not include turnover related to the sales of VAT exempt goods and services, whereas the statistical system is likely to require total turnover.

Another common example is the definition of unemployment. The standard statistical definition[17] is:

> *"The "unemployed" comprise all persons above a specified age who during the reference period were:*
> *(a) "without work", i.e. were not in paid employment or self-employment*
> *(b) "currently available for work", i.e. were available for paid employment or self-employment during the reference period; and*
> *(c) "seeking work", i.e. had taken specific steps in a specified recent period to seek paid employment or self-employment."*

However definitions of unemployment in administrative sources are more often based on the number of people claiming unemployment benefits, or registered as looking for work. Some people who are out of work may not register as unemployed, if they expect to find work quickly, and in some cultures there may be a social stigma attached to claiming unemployment benefits. On the other hand, some people claiming unemployment benefit may not be available for work or actively seeking work, so should not be counted as statistically unemployed.

The first step towards solving the problem of different definitions is to try to understand the differences and quantify the impact. Some differences may have no real impact in practice, so could be safely ignored, others may be systematic, so could be resolved through adjustments to the data. Sometimes it might be possible to derive or estimate the impact of the difference by combining variables from different sources, particularly for financial accounting variables such as the turnover example above. In some cases, it might even be possible to influence the administrative definition.

4.7 Classification Systems

As is the case for variables, the classification systems used within administrative sources may be different to those used in the statistical world. Even if they are the same, they may be applied differently depending on the primary purpose of the administrative source, perhaps focusing on specific attributes of the unit. For example, an administrative source concerned with licensing, health and safety or environmental protection may be more interested in the economic activities of a business that are of most concern to that source, rather than the main economic activity of a business, which is required for statistical purposes.

In other cases, classifications in administrative sources may not be applied at the level of detail required for statistical purposes, or the classification may simply not be a priority variable for the administrative source, resulting in quality deficiencies.

[17] See the Resolution concerning statistics of the economically active population, employment, unemployment and underemployment, adopted by the Thirteenth International Conference of Labour Statisticians (October 1982) http://www.ilo.org/global/statistics-and-databases/standards-and-guidelines/resolutions-adopted-by-international-conferences-of-labour-statisticians/WCMS_087481/lang--en/index.htm

Where classification systems or versions are different, the usual solution is to construct conversion matrices to map the codes in the administrative classification to those in the statistical classification. Such mappings may be one to one, many to one, one to many or many to many. In the latter two cases, some sort of probabilistic allocation may be required.

Box 4.2 – Using a Simple Conversion Matrix

Code 1	Code 2	Weight	
0100	01300	100	**1 to 1 correlation**
0101	01210	26	
0101	01221	14	
0101	01222	29	**1 to many correlation**
0101	25730	11	
0101	74332	20	
0102	03200	100	
0103	01300	36	
0103	74332	64	

This extract from a conversion matrix illustrates the main problems found when converting from one classification system to another. In this case, the codes used in the administrative source (Code 1) are mapped onto those used in the statistical system (Code 2), in a probabilistic way based on weights.

The first issue is therefore how to determine the weights. These can be estimated, but a preferable method, where possible, is to derive them from an analysis of units that have been classified according to both systems, looking at the proportions of units with certain combinations of codes. It may be necessary to constrain these analyses to cover only combinations of codes that are considered valid or plausible to reduce the impact of coding errors.

The first line above shows a one to one correlation, reflected by a weight of 100%. This means that all units with an administrative code of 0100 should be allocated the statistical code 01300. The next five lines show a one to many correlation. If a unit has the administrative code 0101, there are five possible statistical codes. For each of these statistical codes, the likelihood of it being the correct code for the unit is reflected by the weight, thus there is a 26% chance that 01210 is the correct statistical code.

In this case, the probability that a unit with the administrative code 0101 will be given the correct statistical code can be calculated by summing the squares of the probabilities for each combination, e.g.

$0.26^2 + 0.14^2 + 0.29^2 + 0.11^2 + 0.2^2 = 0.2234$

This means that there is a 77.66% chance that a unit with the administrative code 0101 will be given the wrong statistical code. Whilst this likelihood might seem unacceptably high, it should be remembered that even though codes may be wrong at the unit level, providing the weights are accurate, the distribution of units between codes should be correct at the aggregate level, and as long as there are no systematic biases in the application of the conversion matrix, there should be no resulting biases in statistical data for units coded in this way.

It should also be noted that conversion matrices such as the example above are uni-directional. A separate matrix, with different weights, would be required to convert from the statistical codes to the administrative codes. For example, a one to one correlation in one direction may become a one to many correlation in the other direction. This is illustrated in the table above, where there is a one to one correlation between codes 0100 and 01300, but when converting from statistical to administrative codes, 01300 could map to 0100 or 0103.

Where accuracy is required at the micro-data level, it is clear from Box 4.2 that the conversion matrix approach has severe limitations. Various other methods may be possible depending on resources and data availability, but a useful first step is always to gain a detailed understanding of how the classification data are collected and processed by the administrative source, and the nature of the administrative functions they are used for.

In some cases, other variables may be available within the administrative source, which could be used to improve the likelihood of selecting the correct statistical code. One such variable could be the text description from which the administrative code was derived. If this is available, it is potentially of more use to the statistician than the administrative code itself, because the statistician could apply manual or automatic techniques to derive the correct statistical code directly from the description. This method can be used in conjunction with the conversion matrix approach, such that text descriptions are only coded in cases where there is not a one to one correlation between administrative and statistical codes, though there is a risk of potential bias if the quality of coding is different between the administrative and statistical systems.

One approach that has been used successfully in several countries is to develop an automatic coding tool for use in both statistical and administrative systems. This ensures a high degree of consistency of coding, and strongly encourages (but does not necessarily enforce) the use of a common classification system.

In addition to the use of common coding tools, the provision of coding expertise and training to administrative data suppliers can help to improve coding consistency. At the same time, it is always helpful for the statistician to stress the advantages of using a common classification system. It also helps to give early notice of any revisions to the classification system, and to provide as much help as possible to administrative data suppliers during the implementation of the resulting changes.

4.8 Timeliness

There are three separate issues relating to timeliness that affect the usefulness of administrative data for statistical purposes:

- Administrative data may not be available in time to meet statistical needs
- Administrative data may relate to a period that does not coincide with the statistical reference period
- Administrative data may be measured over a period, whilst the statistical requirement is for a specific point in time (or vice-versa).

Considering the first issue, there will generally be some sort of lag between an event happening in the real world, and it being recorded by an administrative source, this is then followed by a further lag before the data are made available to the national statistical organisation. Figure 4.2 below shows the total lag in days between businesses commencing activities and being recorded on the statistical business register in the United Kingdom. Lags relating to births and deaths of enterprises are a major source of business register coverage errors. If these lags are measured, allowance can be made for them in any statistics based on register data.

By analysing lags in this way, it is possible to produce summary statistics to estimate their impact. For example, in the case above, two-thirds of businesses appear on the statistical business register within 2 months of starting activity. The mean lag is around 120 days, but this figure is not particularly useful as it is affected by outliers in the very long tail of the distribution (truncated in Figure 4.2, as the most extreme cases involved lags of up to ten years). Perhaps a more useful measure of average in this case is the median, which is around 40 days. Another interesting feature of this analysis is the small number of negative lags, which can happen when a businesses completes registration formalities well before commencing trading.

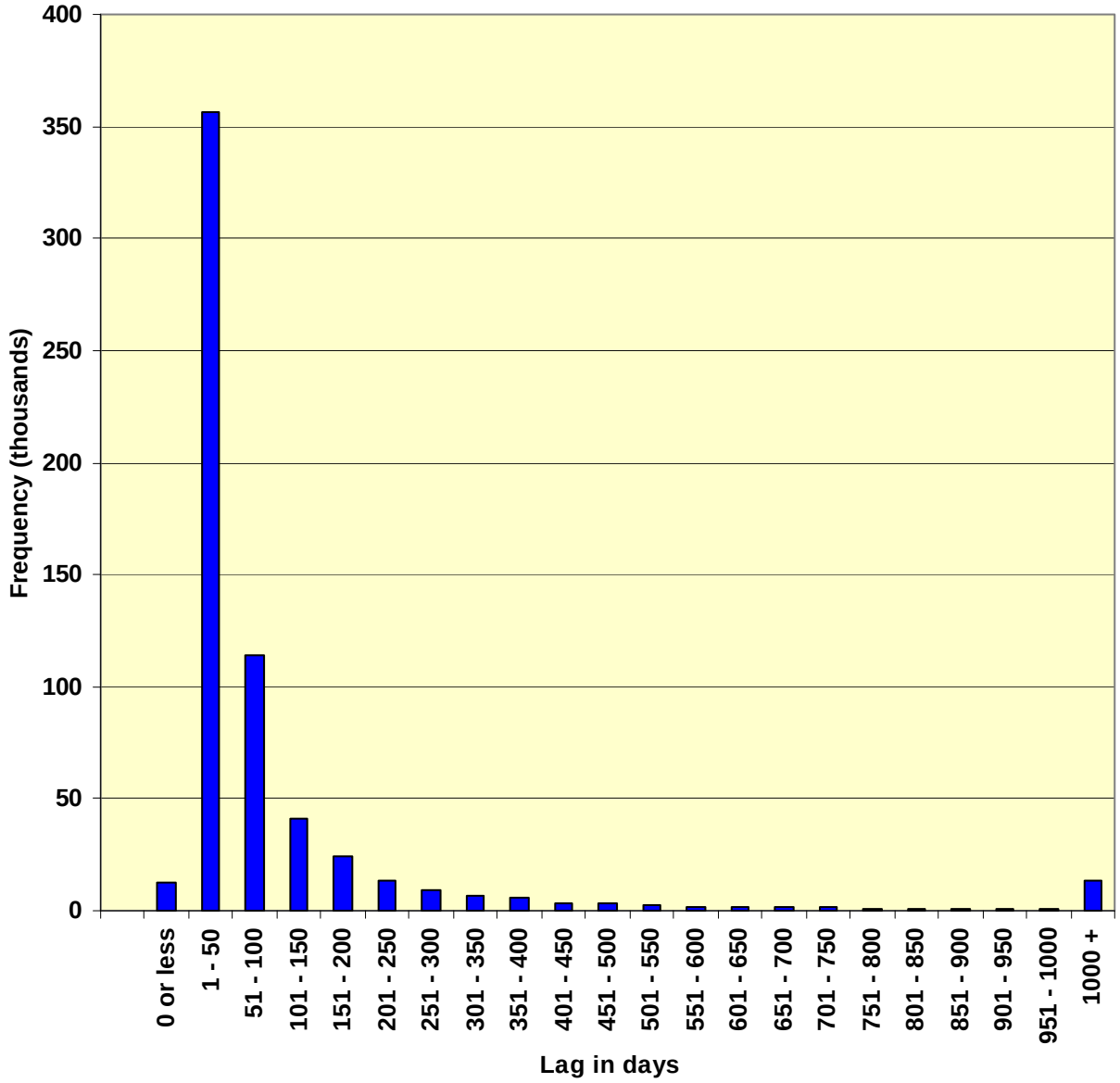
This sort of analysis is clearly important to help the statistician understand the nature and impact of the lags in the sources used to compile statistics. It also gives information that can be used to inform adjustments to improve the quality of the statistical outputs.

The existence and length of lags can make the use of administrative sources difficult for short-period statistics, e.g. a six-month lag would probably be unacceptable for a key monthly economic data series, but would be less of a problem for annual statistics.

The first step to resolving the problem of lags is to understand their impact by preparing analyses such as the one above. Once this has been done, it may be possible to develop models to adjust for their impact[18]. It might also be the case in some relatively stable data series that opposing lags may cancel each other out, for example the business registration lags in Figure 4.2 may be cancelled out by de-registration lags for the purposes of producing data on the business population. It can be dangerous, however, to assume that this is the case, without empirical evidence.

[18] For an example relevant to Figure 4.2, see Annex B of Business start-ups and closures: VAT registrations and de-registrations in 2005 - Guidance and Methodology http://stats.berr.gov.uk/smes/vat/VATGuidance2005.pdf

Figure 4.2 Business Registration Lags in the United Kingdom[19]



When the nature and impact of lags have been determined, it is useful to try to understand what causes them. In some cases, it might be possible to propose changes in the administrative source that would reduce lags. This may be beneficial to both the statistician and the administrative source.

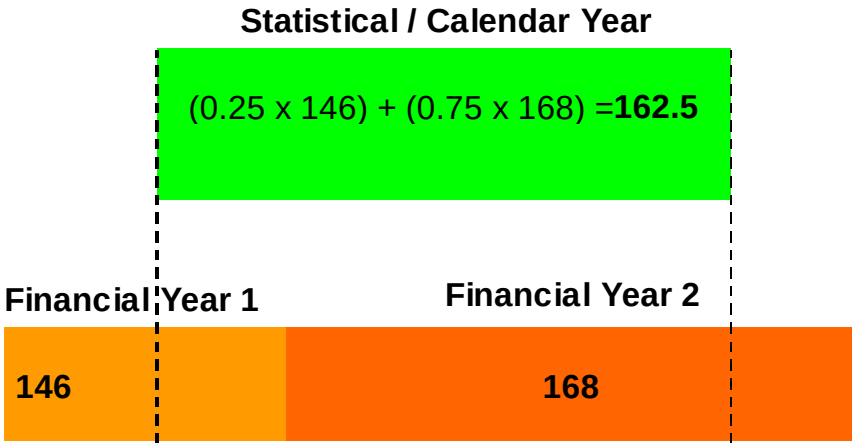
The second issue related to timeliness is that of differing periods, for example data from annual tax returns are often only available several months after the end of the tax year, so are probably not suitable for monthly or quarterly statistics. In some cases, however, annual administrative data can be used for shorter-period statistics, particularly if they are collected on a rolling annual basis. This can happen if there is

[19] Source: Model Quality Report in Business Statistics, Volume III, Eurostat http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/MODEL%20QUALITY%20REPORT%20VOL%203.pdf

a requirement to spread the workload of collecting and processing these data by the administrative source throughout the year. As long as the distribution of the units for which data are collected during the year is sufficiently random, it may be possible to derive meaningful monthly or quarterly statistical trend data from such sources.

Figure 4.3 shows a case where administrative data are based on a financial year running from 1 April to 31 March, whereas the statistical requirement is for calendar year data. The simplest way to convert these data is to add 25% of the value from the first financial year to 75% of the value for the second. This method should give a reasonable approximation if the long-term trend in the data is reasonably stable, though for more volatile series, other, more complex estimation methods may be required.

Figure 4.3 Dealing with Different Time Periods



The third issue concerns the difference between data referring to a specific point of time and data relating to a period (e.g. an annual or monthly average). For example, there may be a statistical requirement for employment data on a specific reference date, whereas administrative data may only give monthly averages.

As in previous examples, the first step is to analyse the impact of the difference, and determine whether it is significant enough to require further action. One possible solution is a model-based mathematical adjustment, e.g. if the statistical reference date is near the start of the month, a model that takes into account the average figure for the previous period may be appropriate. An alternative approach may be to use the results of a relatively small survey to adjust the administrative data.

4.9 Inconsistency between Sources

A specific problem where multiple sources are used concerns inconsistencies between those sources. Data from one source may appear to contradict those from another. This may be due to different definitions or classifications, differences in timing, or simply to an error in one source. This can happen when comparing administrative data with statistical data, or when comparing two administrative (or two statistical) sources.

To resolve such conflicts it is necessary to establish priority rules, by deciding which source is most reliable for a particular variable. Once a priority order of sources has been determined for a variable, it should then be possible to ensure that data from a high priority source are not overwritten by a lower priority source. This process is made much easier if source codes are stored alongside variables for which several sources are available. The use and storage of dates can also be helpful, as even when one source is thought to be more reliable than another, data from that source that are ten years old may not be of higher quality than data for the most recent period from the less reliable source. A simpler method that may be appropriate in some cases is to load data in reverse priority order, allowing data of higher quality to overwrite those of lower quality.

In most cases, there will be several variables of interest, and it is likely that the priorities will differ from variable to variable. For example, an administrative source concerning employment of workers is likely to give reasonable estimates of (legal) employment, as that variable is closely related to the core function of the source. It may not, however be so good for determining the economic activity of the employer, as this may be only of secondary importance for the purpose of the source. Thus if multiple sources are used to derive employment data, it would be necessary to consider the relative quality of each variable in each source in order to derive the optimal statistical data set.

The more data sources that are used, the more complex this comparison process becomes, but having multiple sources often helps to validate data quality. In some cases, certain sources may not be used directly for statistical production, but purely for benchmarking purposes as part of a quality assurance process[20]. The resulting knowledge about the quality of various sources can also be fed back (usually at aggregate rather than unit level, to protect statistical confidentiality) to the source, and can provide a basis for discussions about improving the quality of that source.

[20] An example of benchmarking, using maps to compare the coverage of a statistical business register with that of a commercial telephone directory can be found in the paper “The Development of Small-area Business Statistics in the United Kingdom” at http://live.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf

Box 4.3 – Data From Different Sources

	Source 1: Education Register	**Source 2: Population Register**
Name	Steve Vale	Stephen Vale
Address 1	5 St Peter's St	5 Saint Peters Street
Address 2	Machen	Machen
Address 3	Newport	Caerphilly
Address 4	Gwent	South Wales
Postcode	NP1 8QB	CF83 8QB
Date of birth	28/12/1967	28/12/1997
Occupation	Statistician	Civil Servant
Employer	CSO	Office for National Statistics
Workplace postcode	NP10 9XX	NP10 8XG

This example shows two records containing fictional data about the author (education and population registers do not yet exist in the UK). It is designed to illustrate several common issues when trying to reconcile data from different sources:

- Errors – a simple plausibility check would find the error in the population register, people born in 1997 would still be at school, so could not have an occupation or an employer. As the education register gives the year of birth as 1967, this looks like a simple keying error. Automatic checks can usually find such obvious errors, though have to be used with care, for example, a few genuine cases of children being older than parents have been found in Finland, due to adoptions!
- Timing – the addresses and postcodes given may actually refer to the same building, but at different points in time. The differences could be due to boundary changes between postal areas. This could be determined by consulting historic address files, or by mapping current and historical addresses using geographic information systems.
- Abbreviations – "St" at the end of address line 1 in the education register is a common abbreviation for "Street" so these text strings should be treated as synonyms when they appear at the end of a text line. Note, however, that "St" at the start of address line 1 is used as an abbreviation for "Saint" so again, some care is needed. Similar examples can be found in other languages.
- Timing and abbreviations – in the UK, "CSO" is an abbreviation for "Central Statistical Office" a former name of the "Office for National Statistics", and one that might still be used by those unfamiliar with the change.
- Different spellings – "Steve" and "Stephen" are different variants of the same name, and should be treated as such.
- Classification issues – the occupations "Statistician" and "Civil Servant" are not mutually exclusive. "Statistician" could be said to refer to the profession, whereas "Civil Servant" relates more to the nature of the employment.

- Default values – sometimes when a value is missing, or only partially present, some sort of default value is used. Typical defaults are “Z” or “9999999”. In the UK, when the second part of a postcode was not known, the default “9XX” was often used, as can be seen in the “Workplace postcode” field. Unfortunately the use of this default had to be abandoned when the Post Office started allocating real postcodes ending with “9XX”!

4.10 Missing Data

The problem of missing data is not unique to administrative sources. It can also be due to full or partial non-response to statistical surveys, or even to the removal of data values during the editing process. However, with administrative sources, the issues can sometimes be different, particularly as the problem of missing data can often be more systematic.

The main reasons for this are that a particular variable may not be collected at all by the administrative source, or it may only be collected for certain categories of units where there is a specific administrative requirement. The variable may also simply be a low priority for administrative purposes, so the owners of that source do not see missing data as a problem.

Some of the standard solutions for dealing with non-response in statistical surveys can also be used to solve the problem of missing data in administrative sources. Various imputation methods, such as deductive, ‘hot-deck’ or ‘cold-deck’ imputation are often suitable where the problem only affects some of the units. In cases where most or all of the units are affected, a modelling approach may be more appropriate.

Box 4.4 – Case Study: Dealing With Missing Administrative Data – Turnover per Head Ratios

The two variables most commonly available to measure the size of a business are the number of employees, and the total sales (turnover). However it is common for one or both of these variables to be missing or unreliable for new businesses, particularly smaller ones.

To help resolve this problem, turnover per head ratios can be used to estimate the missing variables. These ratios are constructed using information for similar businesses for which both variables are present and considered reliable, then calculating average turnover per head ratios for different categories based on economic activity and institutional sector.

For example, the following are dummy turnover per head (TPH) values calculated for different classes of the International Standard Industrial Classification (ISIC):

ISIC class	TPH
.....	
45.11	95
45.12	68
45.21	149
.....	

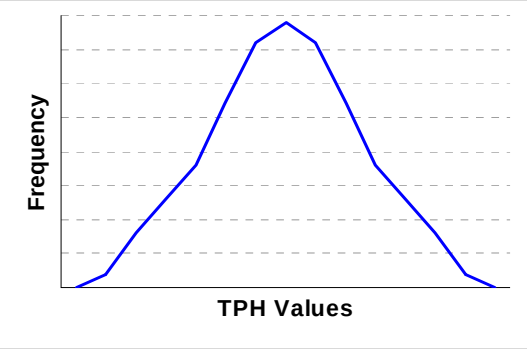
If a business has ISIC class 45.12, and its turnover is 200, but employment is missing, the imputed employment value is:

200 / 68 = 2.94 (rounded to 3)

When calculating turnover per head values, problems with outliers are often encountered, so methods such as trimming (removing the top and or bottom x% of values), and calculating the mean of the inter-quartile range are often used.

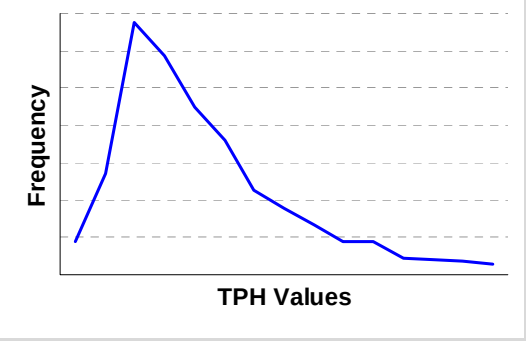
Ratios of this type can also be of more general use for validating updates, matching records from different sources, and detecting errors. For example, by graphing and studying the distribution of turnover per head values, it is also often possible to get useful information about the population of units in question. The following charts are examples of what has been observed from such an exercise:

1. Normal distribution



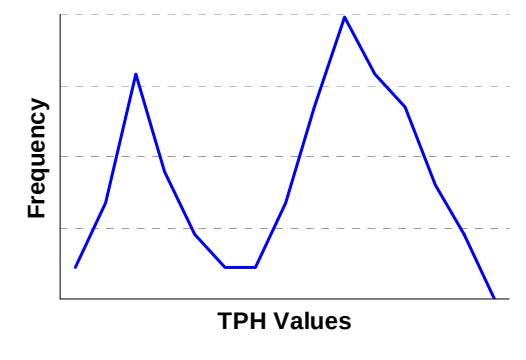
In this case the turnover per head values are distributed evenly around the mean indicating a relative degree of homogeneity amongst the population of units, and very limited impact of outliers.

2. Skewed distribution



In this case there is a clear grouping of units around a relatively low value, but the outliers towards the outer end of the right-hand tail would clearly affect the mean of the distribution. This is a relatively common distribution for turnover per head data, and highlights the need to take measures to reduce the impact of these outliers.

3. Bi-modal distribution



This case illustrates that the population in question is rather heterogeneous, and that it might be worth splitting it into two sub-populations to get more meaningful turnover per head ratios.

4.11 Resistance to Change

One of the main barriers to the more effective use of administrative sources in official statistics, and one of the least recognised, can come from within the organisation. Statisticians may resist the use of administrative data because they do not trust data that they have not collected themselves. They often focus on the negative quality aspects of administrative data, and they have an over-optimistic view of the quality of survey data, often based on the largely untested assumption that survey responses actually comply with statistical norms.

The solution is clearly through better education of statisticians regarding the possibilities offered by administrative sources, encouraging them to take a wider view of all the dimensions of quality, and focus on the impact on data suppliers and users. In this context it is important to determine the real relative quality of survey and administrative data. For example, it is often assumed that data from administrative sources do not meet the requirements of statistical definitions, whereas those from official surveys do. However, there may not be any real difference in practice, particularly if respondents to statistical surveys simply copy values from recent administrative returns, without reading the often lengthy notes about how a particular variable should be defined for statistical purposes.

A further way to help break down the barriers of internal resistance is to show that cost savings from using administrative data do not necessarily mean staff reductions. The resources saved can, at least partly, be used to improve quality or increase the range or frequency of outputs.

4.12 Summary

This chapter clearly shows that there are many problems to overcome when using administrative sources. It also aims to show that others have also faced these problems, and that in most cases it is possible to find full or partial solutions. It can not cover all potential problems the reader may face, particularly those that are source-specific, but the intention is to give ideas that can be adapted to meet specific circumstances.

Overall, it is true to say that most problems encountered in the use of administrative data for statistical purposes, in common with many other areas of statistics, can be overcome, or at least reduced, by effective planning and management, a good knowledge of data sources, creative thinking, and the willingness to exchange experiences and learn from others.

Administrative data often require different processing than statistical sources. Simply substituting administrative data for statistical data without changing the statistical production process will rarely work in practice.

The final thing to remember is that despite all of the problems, the benefits of using administrative data are still often much greater than the costs.

5. Quality and Administrative Data

5.1 Introduction

As mentioned in Chapter 4, concerns about the quality of administrative data are often one of the main barriers to their increased use for statistical purposes. These concerns may or may not be justified, and are often based only on specific aspects of quality, such as timeliness. To properly address these concerns an objective quality management framework is needed; one that considers all relevant aspects of quality, and allows an informed decision to be made.

Many statistical organisations have already put in place some sort of quality framework for data collected via traditional survey methods, but relatively few have extended this approach to cover data from administrative sources[21].

5.2 Defining Quality

The starting point for such a framework is the definition of quality itself. Again, much work has been done in this area by national and international statistical organisations, most of which is based on the international standard ISO 9000/2005[22], which defines quality as:

"***the degree to which a set of inherent characteristics fulfils requirements.***"

Unfortunately this definition is not particularly easy to understand, and needs some further explanation. It can be split into the following parts to aid interpretation:

1) "requirements"

This is usually taken to mean the requirements of the user of specific goods or services, though it could also be argued that the requirements of the producer, or even of society as a whole, should also be taken into account. For example a fast car with a large engine may fully meet the requirements of an individual, but may not meet the requirements of society regarding pollution or road safety. However, the ultimate products of official statistical agencies, the statistics themselves, are usually produced within the public sector as a "public good", so in this case the requirements of these different groups are largely overlapping.

If we consider an administrative data set as a product in its own right though, there can be considerable divergence between the requirements of the producer (e.g. an administrative agency) and the user (a statistical organisation). Furthermore, as the "transaction" is often not on market terms, there may be little incentive for the producer to consider the user requirements. This can result in tensions over quality,

[21] Examples include approaches developed by Statistics Netherlands (http://isi2011.congressplanner.eu/pdfs/950481.pdf) and Statistics Sweden (http://www.scb.se/statistik/_publikationer/OV9999_2011A01_BR_X103BR1102.pdf)
[22] See: http://www.iso.org/iso/catalogue_detail?csnumber=42180

which underlines the requirement for a sound organizational framework as described in Chapter 3.

2) "a set of inherent characteristics"

Users of any goods or services judge quality against a set of criteria concerning different characteristics of those goods or services. This is often done sub-consciously, as in the example of a meal in a restaurant: An individual will judge the quality of that meal in terms of the way the food was cooked and presented, the quantity, the service, the decoration and ambiance of the restaurant, and perhaps several other criteria (for the moment cost is not included, we will return to this later in the chapter). The quality of statistics can similarly be judged against a set of inherent characteristics or criteria.

Several statistical agencies have developed lists of criteria for evaluating quality of statistical data, however the main international agencies have now reached agreement on the following list:

- Relevance - the degree to which statistics meet the needs of current and potential users. Relevance therefore refers to whether the statistics that are needed are produced, and whether the statistics that are produced are needed. It also covers the extent to which the concepts used (definitions, classifications etc.) reflect user needs.
- Accuracy - the closeness of statistical estimates to true values.
- Timeliness - this reflects the length of time between data being made available and the event or phenomenon they describe.
- Punctuality - the time lag between the date that data were actually released and the target (often pre-announced) release date.
- Accessibility - the physical conditions in which users can obtain data: where to go, how to order, delivery time, clear pricing policy, convenient marketing conditions (copyright, etc.), availability of micro or macro data, various formats (paper, files, CD-ROM, Internet…), etc.
- Clarity / interpretability - whether data are accompanied by sufficient and appropriate metadata, whether illustrations such as graphs and maps add value to the presentation of the data, and whether information on data quality is available.
- Coherence / consistency - data from different sources, and in particular from statistical surveys of a different nature and/or frequency, may not be completely coherent in that they may be based on different approaches, classifications and methodologies. They may not, therefore, convey a completely coherent message to users, e.g. users may be confused if two different measures of the same variable are published with different values.
- Comparability - the extent to which differences between statistics are attributed to differences between the true values of the statistical characteristic, or to methodological differences. Comparability includes:

- Comparability over time – the extent to which data from different points in time can be compared.
- Comparability through space – the extent to which data from different countries and/or regions can be compared.
- Comparability between domains – the extent to which data from different statistical domains can be compared.

This list of criteria can be used in two ways relating to administrative data. Firstly it can be used to assess the quality of the resulting statistics, and to compare data based on administrative sources with those based on surveys. Secondly, the list can be used to help evaluate the quality of different administrative sources themselves[23]. For example, if a statistician is fortunate enough to be faced with a choice of two or more administrative sources, it can help to determine which source has the higher quality.

However, if the list is being used to assess the quality of administrative data, it should be noted that absolute accuracy can be difficult to determine if there is not sufficient supporting information about the population and the collection process. In this case, two factors should be considered, the credibility of the source and the plausibility of the data, i.e. whether the source is trusted, and whether the data look reasonable when compared to other sources, and to the values the statistician would expect. For a more objective measure, some sort of quality survey may be needed to determine the correct values of certain variables.

The closeness of administrative units and variable to the units and variables required for statistical purposes can be an important factor in determining the quality of an administrative source. The fewer transformations required, the lower the risk of error or bias. This aspect can be considered as part of the criterion of coherence.

5.3 The Constraint of Cost

Cost is deliberately excluded from most lists of statistical quality criteria, as it is considered to be more of a constraint. Once quality has been determined, cost is added to the equation to allow practical decisions on cost-efficiency to be made.

Cost is, however, particularly important in the case of administrative sources, because where they are shown to deliver a lower absolute level of quality than survey data, they may still have a sufficient cost advantage, which could make them the most cost-efficient option. It may also be possible to channel some of the cost savings into improving quality, thus reducing or eliminating the quality gap.

[23] For an application and extension of this approach, see the discussion paper from Statistics Netherlands “Checklist for the Quality Evaluation of Administrative Data Sources”: http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf

5.4 Quality Measurement in Practice

To fully understand the quality of administrative sources, and their impact on the quality of statistics, we need to consider three elements:

1) The quality of incoming data

The incoming data, whether they are from administrative or survey sources, can be judged against set of criteria such as those listed above. The most important criteria are likely to be timeliness, and relevance in terms of the extent to which the coverage and concepts of the source meet requirements. Comparability with other sources can also be important, and some sort of exercise to reconcile data from different sources may be necessary from time to time to get a clear picture of quality. Quality check surveys are sometimes used for this purpose.

One point worth bearing in mind is the extent to which the data subject has an interest in the quality of the data. The amount of effort and care put into providing the data will vary according to the perceived value or importance of the data collection, thus data subjects may, in some cases, provide better quality data for administrative purposes than they do for statistical purposes.

2) The quality of data processing

Even if the incoming data are perfect, their quality can still be affected by the different processes they go through before they are used for statistical outputs. Ideally processing should improve quality, but unfortunately this is not always the case. Examples of how data processing can affect quality include:

- Data matching and linking – too many false matches will lead to errors in the data; too many false non-matches will lead to duplication, which will overstate the size of the population of interest, and possibly introduce bias.
- Outlier detection and treatment – using outlier detection methods to detect errors can help to improve the quality of the data, and generally the more extreme the outlier, the more likely it is to be an error. However over-zealous treatment of outliers will result in genuine data values being altered and can lead to important trends in the data being missed.
- Quality of data editing – as for outlier detection and treatment, data editing should improve quality, but if not done carefully it can introduce error and bias[24].
- Quality of imputation – if imputation is used to fill missing values or records it can help to improve coverage, but again the methods used need careful scrutiny to avoid the introduction of bias.

One very important principle that should always be followed, particularly when processing data from administrative sources, is to keep a copy of the raw data (and any associated metadata) to refer back to if necessary. Comparisons of data before

[24] For a comprehensive collection of papers on different data editing issues, see the working papers of the Statistical Data Editing Work Sessions organised by the UNECE - http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing

and after processing can help to assess the quality of that processing, and to identify any specific problems.

3) The quality of statistical outputs

The usual interpretation of the ISO quality definition by statistical agencies is that quality is all about meeting user requirements. The quality of statistical outputs is therefore determined in this context. This means that it is necessary to determine these requirements, to discuss them with users, and to get regular feedback, for example via user satisfaction surveys.

Moving from survey to administrative sources will clearly have an impact on output quality. Typically this impact may be positive for some quality criteria, and negative for others. In all cases, it is necessary to get an overall view of the impact, giving greater weight to those criteria the users consider to be the most important. For example, users may feel that an improvement in timeliness more than compensates for a reduction in accuracy, particularly for short-period economic data. Another consideration should be the impact on time-series data, and whether it is possible to construct a consistent series of sufficient length following the change.

It can be particularly important to give at least as much weight to the views of users as to the perceptions of statisticians, which may, in some cases be too heavily focussed on traditional notions of accuracy. Overall, it is vital that any judgement of the impact on statistical outputs is based on objective evidence rather than on supposition, as this is the only way to counter the potential for resistance to change as described in Chapter 4. One way to ensure this is to use quality reports, following standard templates[25], to document and communicate the impact of changing data sources.

5.5 The Role of Metadata

Metadata[26] are vital for informing both producers and users about data quality. They should be present at all three of the stages referred to in the previous section. Incoming data should be accompanied by sufficient metadata to fully understand them, and to ensure that values are correctly allocated to the relevant variables. Detailed documentation on the concepts, definitions and purpose of the source, as well as on the collection and processing methods used, is also important. This will give a better understanding of potential quality issues, and should form the basis for data editing rules in the processing stage.

During data processing it is important to record what has been done to which records and values. This not only provides vital information for assessments of processing quality, but also provides a mechanism to investigate any potential problems in the process and undo any errors.

[25] For example those proposed by Eurostat, see: http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

[26] Data that define and describe other data (Source: ISO/IEC FDIS 11179-1 "Information technology - Metadata registries – Part 1: Framework", March 2004)

Statistical outputs should be accompanied by sufficient metadata to allow users to retrieve them, interpret them correctly, and form an opinion on their quality. For regular and heavy users of the outputs, full documentation of all three stages, preferably following a standard format, will provide the necessary information to enable them to draw the correct conclusions from the data. Communication of quality can often be difficult to get right, as some users want full details, whereas others are happier with very high-level summary indicators. A metadata model that allows users to see different levels of information, starting with a summary, but with an option to see greater detail, is perhaps the most appropriate.

5.6 Summary

The best way to assess the quality of an administrative source is to build up a thorough knowledge of that source, including the primary purpose of the source and the way the data are collected and processed. Thorough understanding of a source will allow a more accurate assessment of strengths and weaknesses.

To assess the impact of using different sources, it is necessary to combine knowledge of the sources and the processes used to convert them to statistical outputs with the views of the users of those outputs. This will then allow an objective and holistic assessment of the impact of using administrative data versus statistical survey data.

6. Data Linkage and Matching

6.1 Introduction

There are two main approaches to the use of administrative data in the statistical production process:

- As a direct source for statistics
- As an indirect source, in combination with other sources

If data from several administrative sources are used to supplement survey data or populate a statistical register, the national statistical organisation will need to find some way of linking those data. This will typically take the form of matching, which can be defined as the linkage of data from different sources based on common features present in those sources.

6.2 Common Identifiers?

If these common features include some sort of common reference or identification number (referred to as a common identifier from this point onwards), the process can be referred to as exact matching, and is relatively easy. In exact matching there are two possible outcomes, either two records from different sources match exactly on the basis of the common identifiers used, or they don't. In other words, a record with the identifier 123456 will match to a record in a different source with the same identifier (assuming the sources are covering the same units!), whereas it will not match to a unit with the identifier 123457.

Exact matching depends heavily on the quality of the matching variables used in each source. If there are errors in the common identifiers in at least one source, there is a high risk of either matching the wrong units, or failing to match units that should be matched. For this reason, even when common identifiers exist in all the files to be matched, it may not be sufficient to rely on exact matching alone.

Sometimes identifiers can include check digits, i.e. one or more characters that are generated according to a standard algorithm based on the other digits in the identifier. If check digits are present, they should help to guarantee a certain level of quality by eliminating most typing or reading errors.

6.3 Matching Keys and the Concept of Distinguishing Power

Where common identifiers are not present, or are not of sufficient quality to give the required level of accuracy in matching, it is necessary to consider using other variables common to the sources involved. The variables chosen are often referred to as "matching keys". Note: it is not always necessary for these variables to be present in both sources, as in some cases they can be derived (for example, see the discussion on turnover per head ratios in Box 4.4). When variables other than

common identifiers are used, the matching routines tend to rely on probabilities to determine which records match.

The variables most commonly used for this sort of probabilistic matching are name, address, date of birth, occupation or economic activity code. The choice of variables to be used for matching should take into account both the “distinguishing power” of each variable. Distinguishing power relates to the uniqueness of the values of the matching key. Some variables have higher distinguishing powers than others:

- High distinguishing power: reference number, full name, full address
- Low distinguishing power: sex, age, city, nationality

Within a variable such as “full name”, it is also possible for some values to have a higher distinguishing power than others. Names that are unique will have the highest distinguishing power, whereas those that are more common (e.g. John Smith in many English-speaking countries) will have a much lower distinguishing power.

Distinguishing power can also depend on the level of detail, e.g.

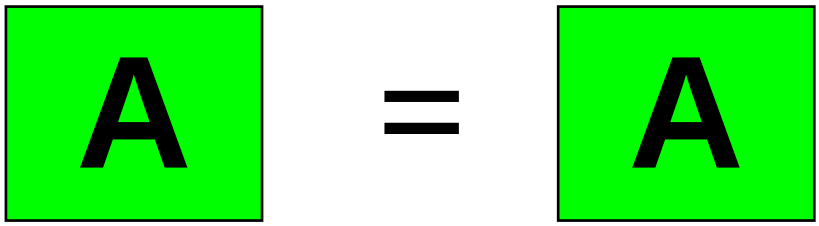
- “Born 1960, Paris” = low distinguishing power
- “Born 23 June 1960, rue de l’Eglise, Montmartre, Paris” = high distinguishing power

Therefore careful choice of matching keys, taking account of the concept of distinguishing power, can have a significant impact on the success of a matching exercise.

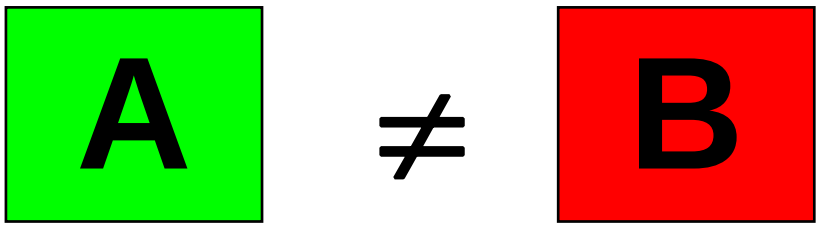
6.4 Some Basic Matching Terminology

When two records are compared, they can be referred to as a “pair”. The following scenarios illustrate the main potential outcomes from applying matching techniques to that pair of records:

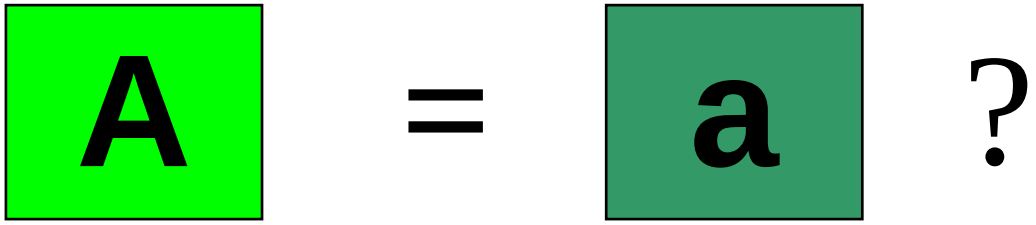
1) Match - A pair that represents the same entity in reality



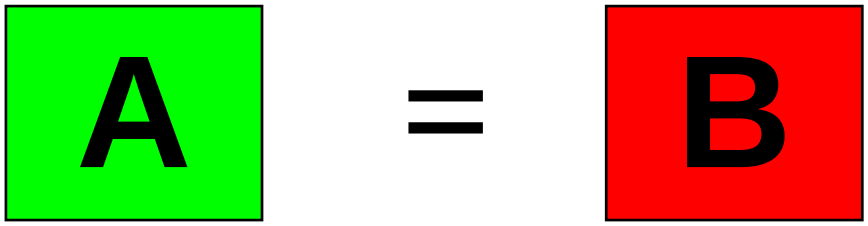
2) Non-match - A pair that represents two different entities in reality



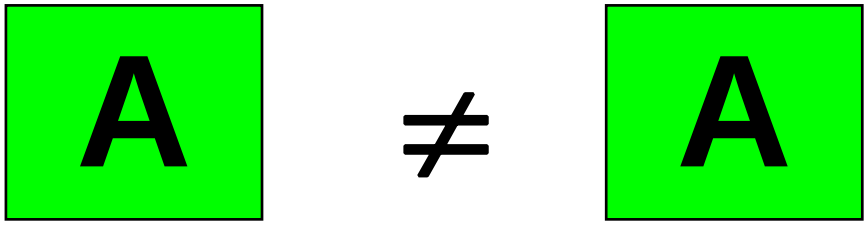
3) Possible Match - A pair for which there is not enough information to determine whether it is a match or a non-match



4) False Match - A pair wrongly designated as a match in the matching process (false positive)



5) False Non-match - A pair which is a match in reality, but is designated as a non-match in the matching process (false negative)



To get a better appreciation of matching concepts and issues in practice, please see the matching exercise at the end of this chapter. It uses made-up, but realistic data to illustrate how matching without common identifiers requires a certain amount of judgement, and how matching can often be more of an art than an exact science. Any form of probabilistic matching is likely to result in a certain proportion of false

matches and false non-matches, as well as the need for further investigation of possible matches.

6.5 Matching Techniques

Matching techniques can be split into two basic categories:

1) Clerical matching – by definition this requires significant human input, so is therefore likely to be:

- **Expensive**
- **Inconsistent**
- **Slow**
- **But; intelligent**

2) Automatic matching – once operational (i.e. ignoring one-off set-up costs), this approach minimises human intervention, so is likely to be:

- **Cheap**
- **Consistent**
- **Quick**
- **But; of limited intelligence**

The best solution is therefore to use an automatic matching tool to find the obvious matches and no-matches, and to refer possible matches to specialist clerical staff. To be cost-efficient, the aim must be to maximise automatic matching rates whilst minimising clerical intervention. The remainder of this chapter considers the main features of automatic matching, and how it can be used and improved in practice.

6.6 How Automatic Matching Works

Automatic matching tools usually follow a similar sequence of steps, though depending on the particular application, some steps may be omitted or others may be added. The most common steps are:

1) Standardisation

This step is mainly used for text variables, or variables that should conform to a specific format. Examples of standardisation processes are:

- Abbreviations and common terms are replaced with standard text, for example the text string “ltd” could be converted to “limited”, and “mfg” to “manufacturing”.
- Common variations of names are standardised, for example there may be different versions of the name of a city (“Brussel” / “Bruxelles” in Belgium, “Derry” / “Londonderry” in Northern Ireland). A similar process is needed for person names where there are different spellings of the same name (“Jane” /

"Jayne") or common short versions of a name ("Bill" / "William"). This is a similar process to, and can possibly be combined with the standardisation of abbreviations.

- "Noise" words are removed – typically these are words or phrases with very low distinguishing power, examples could include "road" or "street" in addresses.
- Postal codes, dates of birth etc. are given a common format, for example "3 January 1985" could be converted to "030185".

The process of standardisation is heavily language dependent, and may also vary according to the type of records being matched, thus the above examples only illustrate the process. Each instance of matching will require prior work, usually based on an investigation of the data, to determine which standardisation rules should be applied.

Standardisation can also be seen as a form of data cleaning, and as such, carries a risk that it could distort or reduce the quality of the data, and even in extreme cases reduce the likelihood of finding a correct match. Such risks are usually very small, and are usually due to ambiguity in the string being standardised. Examples in the English language include the abbreviation "St." which could refer to either "street" or "saint", and the name "Chris", which could be a short form of "Christopher" (male) or "Christine" (female).

Another type of standardisation sometimes used as an initial step in a matching process is to check addresses against a definitive list, usually from the national postal authority. This can range from a check that the combination of postal code and town / city / region is valid, to a full check of the entire address. The success of such a check will obviously be heavily dependent on the quality of the reference file of addresses used.

If the result is that a "cleaned" address is used, it is good practice to also keep a copy of the raw data. In several cases (including matching business data in the UK), it has been found that using cleaned addresses increases the likelihood of matching some records, but decreases it for others. Combining the results of two parallel matching exercises, one using cleaned addresses, and the other using the raw versions can often give the best results.

Two other potential consequences of using cleaned addresses should also be noted, even though they are not strictly related to matching. The first is that in some countries, postal authorities may give a discount for bulk mailing where the addresses used conform to certain standards, so this may help to offset the costs of the cleaning and matching process. On the other hand, substituting a cleaned address for that supplied by a respondent may in some cases cause annoyance to the respondent. If cleaned addresses are used for mailing statistical questionnaires, this may affect response rates. These are further arguments for storing both cleaned and raw data whenever possible.

2) Parsing

Parsing can, to some extent, be seen as an extension to standardisation. In this step, text is converted from a form that is readily recognisable by humans, to a form that is more logical for computer processing, and therefore more likely to correctly match. The resulting text strings are often referred to as matching keys. Early approaches to parsing in the English language often used the “Soundex algorithm”, first patented in 1918. This algorithm, or derivations from it, form the basis of many matching applications. However, parsing rules vary considerably between languages, and should be tuned to give the best results for the data concerned.

Examples of parsing rules could include the following:

- Converting letters or groups of letters with similar sounds to a common string, e.g. “f”, “v” and “ph” to “f”.
- Removing silent letters, e.g. the “h” in the name “Thomas”.
- Converting all characters to either upper or lower case.
- Converting vowels to a single character.
- Removing vowels at the end of a name or word.
- Replacing double letters with single letters, e.g. “Ann” becomes “An”

For example, during parsing using all of the above rules, the string “Steven Thomas Vale” could be converted to “stafan tamas fal”. The string “Stephen Tomos Vael” would also give the same result, showing how parsing can help improve matching rates by reducing the impacts of different ways of spelling a name, and of spelling errors. It should also be noted that varying the order in which the parsing rules are applied could affect the outcome.

As with standardization, however, if parsing is not sufficiently well adapted to the data to be matched, there is a risk that it could do more harm than good. At least in the initial stages, the impact of the parsing routines should be carefully analysed, and in all cases, a copy of the raw data should be retained for comparison purposes.

3) Blocking

If the file to be matched against is very large, it may be necessary to break it down into smaller “blocks” to save processing time. There are several ways to do this, for example, if the record to be matched has an address in a certain town, it may only be necessary to match it against the block containing other records from that town, rather than all records for the whole country.

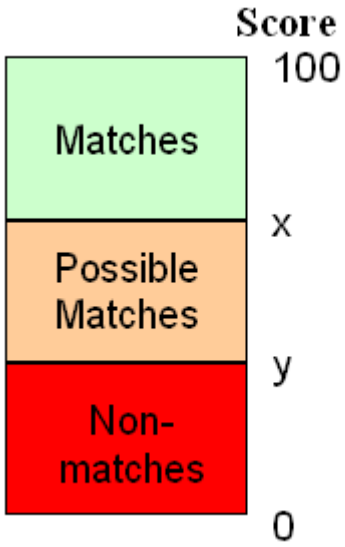
Blocking must be used with care, and is often likely to result in reductions in the overall matching rate. However, if these reductions are minimal, and the gain from faster processing times is substantial, blocking can improve the cost-efficiency of the matching process. In some cases it may even be appropriate to have two or more attempts at matching, using different blocking criteria. For example, after applying relatively restrictive matching criteria to the full dataset, it might prove advantageous to re-process the sub-set of records that do not initially match, using successively less restrictive blocking criteria.

Blocking is clearly most appropriate for very large datasets, such as individual records from a population census, but as a matching technique, it is likely to decline in value as computer power and processing speed increases.

4) Scoring

Most automatic matching routines use some form of scoring to assess the likelihood of a match between two records. Scores are allocated based on how closely the matching variables agree. These scores can be used to determine whether a pair of records is considered to be a definite match, a possible match or a non-match. Figure 5.1 shows how the different categories can be assigned based on threshold scores x and y, in this case expressed as scores out of 100.

Figure 5.1 The Use of Threshold Scores to Determine Category of Match



The next logical question is to how to determine the values of x and y. One option is to use a model based approach, as proposed by Fellegi and Sunter[27], though in practice, a trial and error method is equally likely to be used.

For repeated matching exercises, data quality will vary over time, so a periodic re-assessment of the values of x and y is needed. Similarly, changes in the requirements for the matched data, or in the resources available for matching can lead to revised thresholds. Thresholds may also vary significantly between different datasets.

In setting the values of x and y, it is also necessary to consider the impact of different types of matching error. If a false match is likely to lead to disclosure of statistical information about one unit to another, then the value for x should be set sufficiently high to make the risk of a false match acceptably low. However, if there is no risk of disclosure, and the results of the matching will be used in a study where a certain proportion of false matches is unlikely to have a significant impact on the results, the value of x can be lower.

The availability of clerical resources to investigate possible matches will often in practice place a constraint on the distance between x and y. In all such cases, clerical intervention should be prioritised. This may be by score, so that those possible matches with the highest score are checked first, as this could be assumed to give the most benefit. Alternatively, some other characteristic of the units involved (e.g. number of employees for businesses) can prioritise the clerical work so that it minimises the impact of potential duplication.

[27] See A Theory for Record Linkage, Ivan P. Fellegi and Alan B. Sunter, http://www.jstor.org/stable/view/2286061

6.7 Matching Applications in Practice

Although they often work in rather different ways to that described above, the matching applications most familiar to many people are Internet search engines. They take a text string (typed in by the user) and then search for web pages related to that string, often scoring the results and returning them in order of perceived relevance. Some form of parsing may also be apparent in the results, or through the suggestion of alternative spellings.

Internet search engines also provide a good demonstration of the concept of distinguishing power. For example, at the time of writing, a search on www.google.com for the text string “matching” returned around 700 million results, whereas “statistical matching” returned about 30 million results, and “parsing techniques in statistical matching” returned around 1.6 million results. More detail clearly helps to focus the search.

In the world of official statistics, there have been two main approaches to developing data matching applications:

- Using “off the shelf” commercial software, e.g. Informatica Identity Resolution (incorporating SSAName3)[28]. It should be noted however, that some form of customization is likely to be needed before any commercial package can be used to its full potential.
- Developing matching routines in house, e.g. software developed by the US Census Bureau[29], Statistics Canada[30] and ISTAT, the Italian statistical office[31].

An alternative approach to matching is the “trigram” method, which works by splitting text strings into groups of three characters, and then calculating the proportion of identical groups between two strings.

For example, matching the string “Steven Vale”:

Ste/tev/eve/ven/en /n V/ Va/Val/ale

To the string “Stephen Vale”:

Ste/tep/eph/phe/hen/**en /n V/ Va/Val/ale**

Results in six matching trigrams (shown in bold), out of a total of thirteen unique trigrams from both strings, thus giving a score of 6/13 or 0.46. Parsing of the strings may help to improve the score, but as discussed above, may also introduce errors[32].

[28] http://www.informatica.com/products_services/identity_resolution/Pages/index.aspx
[29] http://www.census.gov/srd/papers/pdf/rr2001-03.pdf
[30] http://www1.unece.org/stat/platform/display/msis/G-Link
[31] http://forge.osor.eu/projects/relais/
[32] A practical application of this method, programmed as SAS code, was demonstrated by Statistics Finland within a Eurostat project to develop statistics on business demography.

Box 6.1 – Case Study – Extracts from "Matching Records Without a Common Identifier - The UK Experience" by Steven Vale and Mike Villars

This text is derived from the full paper, which can be found at: http://www1.unece.org/stat/platform/download/attachments/56230020/matching+paper.pdf?version=1

The UK statistical business register uses data from several administrative and statistical sources, the most important of which are Value Added Tax records and Pay As You Earn income tax records. There is a considerable overlap in the coverage of these two sources, so to minimise duplication it is essential to check that new units from each source are genuine, and haven't already been added from the other source. Each source has its own system of unit identifiers, which means that matching based on names and addresses is the best solution.

Input files are processed in four phases:

- Cleaning - This routine edits the name string, removing special characters and replacing lower-case with upper-case.
- Formatting - This routine edits the name string into separate words, removing "stop words"; replacing selected words and concatenating prefix words.
- Standardisation - This routine "standardises" the name, for example removing double characters.
- Key generation - This generates codes based on the input text, e.g. if the input is "Steven Vale" the keys produced are:

  STEVEN → STAFAN → XJXM$$$; and VALE → VAL → YLVO$$$$

YLVO$$$$ is the key for the last part of the name, and is used as the major key. It is checked against a table of namekeys generated from the names of each record held on the register, to find potential matches. The input name, address and postcode are compared with the name, address and postcode of each of these potential matches and given a score out of 100. If the score is >79 then the pair is considered a definite match. If the score is between 60 and 79 then it is a possible match. Any lower score is regarded as a non-match.

Duplicates on the definite match list are removed, as well as records on the possible match list that also appear on the definite match list. The records on the definite match list are then linked automatically to their corresponding units on the register. The records on the possible match list, and larger non-match records are reported for clerical checking. For a typical update, around 37% of records are definite matches and 35% are possible matches (of which approx. 80% can be matched clerically).

One problem encountered was the use of "Trading as" or "T/A" in names e.g. "Mike Villars T/A Mike's Coffee Bar". In this case, "Bar" would be used as the major key, but has a low distinguishing power as there are many bars in the UK. The solution was to split the name so that the last word prior to "T/A" i.e. Villars is the major key.

Annex to Chapter 6 - Matching Exercise

This exercise contains five examples where a new record has been automatically matched against an existing set of records. No definite matches have been found, but the five highest scoring possible matches are presented for clerical checking. These data are realistic, but are not actually real. Please choose the best match for the new record. Alternatively, if none of the possible matches seem close enough, you can decide that there is no match. Answers are given after Example 5.

Example 1

New record		Possible matches	
Name:	Bob the Butcher	1	Bob Daley Butchers 17 Barwick Green Sidcup Kent DA15 8HP
Address: Postcode:	16 "Lawrence Street Southfleet Gravesend DA11 7ZP	2	Brian Dunn Brians Family Butchers 16 Pembroke Close Pembroke Street Dover Kent DA6 1FB
		3	Mr B Dunn and Mrs V Dunn Brian's Family Butcher Pembroke Street Gravesend Kent DA6 1AA
		4	B & B Butchers Mr B Jones 3 Clive Road Dartford Kent DA1 5RH
		5	B Washbrook Bob the Butcher 16 Lawrence Drive Castle Lane Southfleet Gravesend Kent DA11 7ZF

Example 2

New record		Possible matches	
Name:	Cars of Southfleet	1	Fleet Motors 31-35 Old Dover Road Dartford Kent DA15 7JF
Address:	3-5 Old Hill Southfleet Dartford	2	Southwold Cars 1A Southwold Close Greenhithe Kent DA23 9BC
Postcode:	DA1 9KT	3	Mr D Crane T/A Southeast Cars 12A Old South Road Greenhithe Gravesend Kent DA2 9BN
		4	Mr C James & Mr G Smith Fleet Motors 29-35 Old Dover Road Fleet Kent DA15 9XX
		5	Southfleet Cars 33 Old Hill Southfleet Dartford Kent DA1 9XT

Example 3

New record		Possible matches	
Name:	Retail Co-operative Limited	1	Mr A Cooper Paintcraft Unit 132 Greenway Estate Lower Station Lane Welling Kent DA18 6GT
Address:	35, Station Parade Station Road Dartford	2	Retail Co-op Ltd 030001 35 Station Street Dartford DA1 7DH
Postcode:	DA1 7ED	3	Co-operative Funeral Services 362 Longfield Street Dartford DA1 1HD
		4	Co-operative Funeral Services Ltd, CFS (No14) Ltd & CFS Pension Fund 29 Station Street Bexleyheath Kent DA32 4RH
		5	Arts Co-operative 62 Highfield Street Dartford DA21 8JD

Example 4

New record		Possible matches	
Name:	Dr James Johnson	1	Mr James John Cunningham 35 Griffin Drive Darenth Dartford Kent DA4 6FF
Address: Postcode:	Griffons Penny Lane Eynsford Dartford DA46 8FF	2	Mr John Jameson 56 Whinfell Road Gravesend Kent DA21 8GF
		3	Mr James Johnson 123 Penny Lane Aynsford Kent DA46 3JF
		4	John James 23 Perry Lane Dartford Kent DA28 3PF
		5	Mr James John Smith 18 Cornfield Lane Eynsford Dartford Kent DA46 8FF

Example 5

New record		Possible matches	
Name:	Redipure Ltd	1	Redipure Limited Perseverence House 36A Cross Road Howley Dartford Kent DA27 8RR
Address: Postcode:	26A Queens Rd Welling DA13 8RS	2	Eradicure Ltd Perseverence House Cross Rd Howley Dartford Kent DA27 8RT
		3	Redpull Ltd 152 Lower Wickham Lane Wellington Kent DA13 8ED
		4	Redpull Ltd 12 Lower Wickham Welling Kent DA13 3ED
		5	Redipure Holdings Ltd Crossroads Howley Dartford DA12 3LF

Answers

This exercise shows that there is rarely 100% certainty in matching. The answers below reflect the greatest likelihood of a match according to clerical matching experts.

Example 1 – The most likely match is with existing record number 5. The trading style of this record matches the name of our new record and the addresses are fairly similar. There is one character different in the postcode, “P” instead of “F”, which could easily be a transcription error in one of the records.

Example 2 – Again, the most likely match is with existing record number 5. The names and addresses are sufficiently similar, and as in example 1, there is only one character different in the postcode. This case also highlights an interesting issue, in that the existing records 1 and 4 may also be a match. This could indicate duplication amongst existing records, and shows the value of matching a dataset against itself periodically to reduce the risk of such duplication.

Example 3 – The closest match is with existing record number 2. The main differences concern the use of abbreviations in the name of the existing record (Ltd – Limited; Co-op – Co-operative). This suggests that the automatic matching routine is not sufficiently good at relating abbreviations to their full versions. Abbreviations such as these are often specific to a language or even to a data set, and show the value of being able to customize automatic matching tools according to the types of data being matched.

Example 4 – The closest match is with existing record number 3. Record number 5 is an exact match for most of the address, as well as the postcode, and would potentially score higher in automatic matching. This illustrates the risk of setting the positive match threshold too low.

Example 5 – Based purely on the evidence here, there does not seem to be a match. However, this case illustrates the value of using additional knowledge in clerical matching. The abbreviation “Ltd” in the name of the new record indicates that it is a limited liability company. In many countries, limited liability companies must, by law, have unique names. This suggests that if the matching is intended to link units in the same enterprise, the new record should be linked to existing record number 1. The different addresses may simply refer to different sites (local units or establishments) that this company operates from. Improving the automatic matching routine to recognise corporate businesses, and put a much greater emphasis on the name in such cases, could therefore help to improve the automatic matching rate. This strategy was successfully adopted in the matching routines used for the UK statistical business register.

7. 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."*[33]

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[34].

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

[33] "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.
[34] See the Fundamental Principles of Official Statistics (Principle 6) http://www.unece.org/stats/archive/docs.fp.e.htm

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[35] use data on legal form, economic activity classification and foreign ownership in their statistical business registers to derive the institutional sector[36] 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

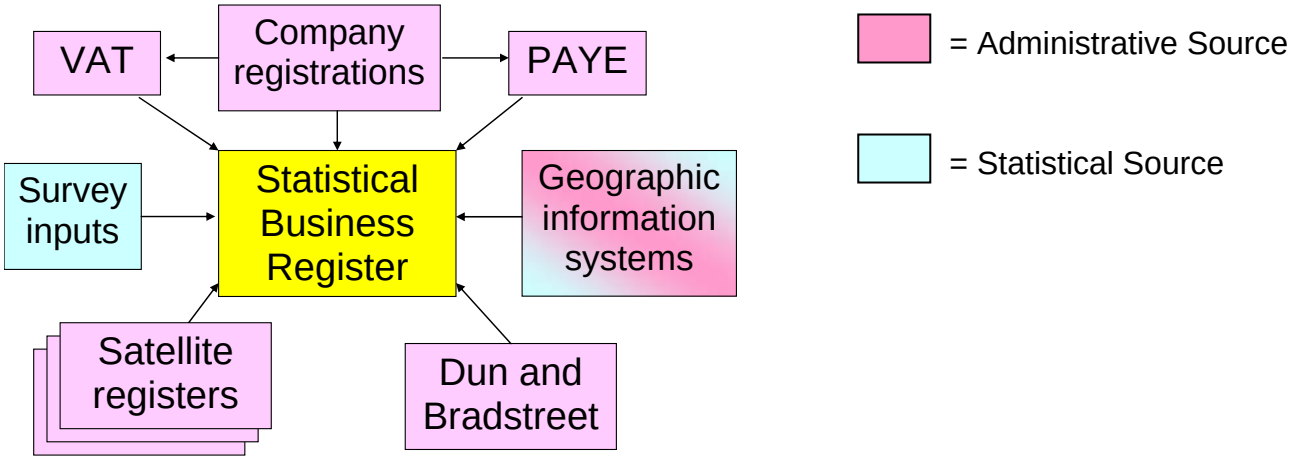
Figure 7.1 below is a simplified model of the sources used to maintain the statistical business register in the United Kingdom. It deliberately shows the statistical register at the centre, as the tool to combine and reconcile the data from the various sources. It also introduces the concept of satellite registers, which will be discussed in detail later in this chapter, and the idea that sources may already be a mixture of

[35] 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.

[36] See "System of National Accounts 2008", Chapter 4 - http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf

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)[37] 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.

[37] See: http://www.abr.gov.au

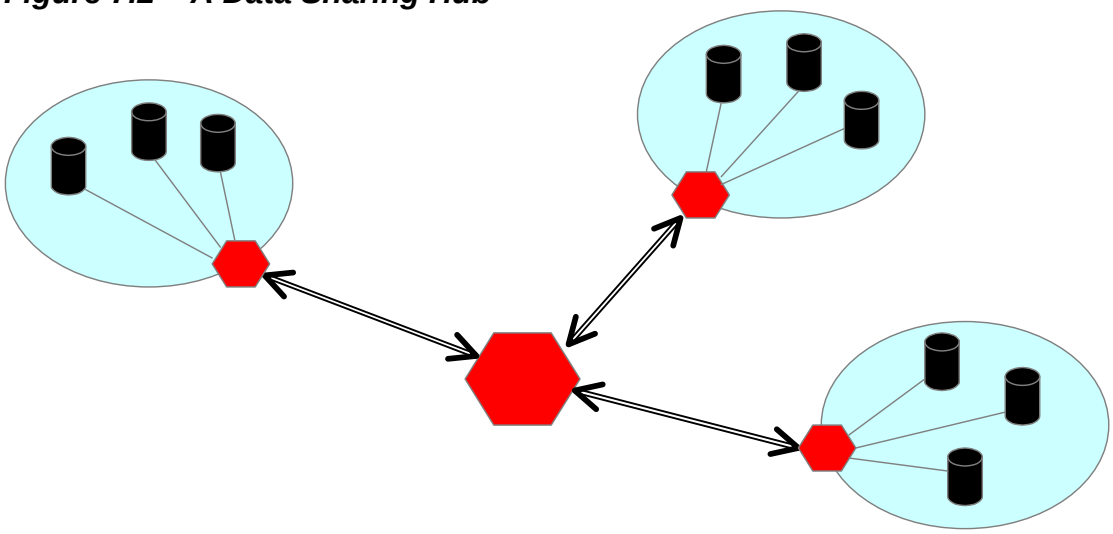
The result is that the ABR is a suitable basis for the statistical business register, for all but the largest and most complex businesses. In fact the statistical business register has a clear two-tier approach[38]. Most records are direct copies from the ABR, and are only maintained from that source, 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[39]. 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



[38] For more information see: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8165.0Explanatory%20Notes1Jun%202007%20to%20Jun%202009?OpenDocument

[39] For more information see: http://www.unece.org/stats/documents/ces/sem.46/5.e.pdf

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"[40]. 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

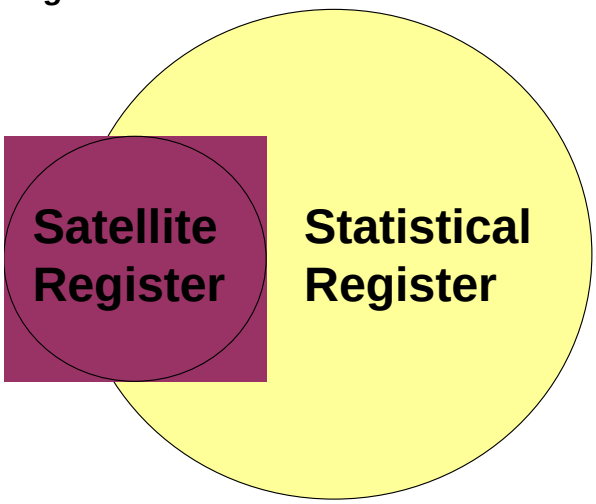


Figure 7.3 shows how a satellite register relates to a statistical register. This diagram can be interpreted both in terms of units covered and variables contained. In both cases there is a degree of overlap, but the satellite register also brings additional information, either additional units, or additional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

[40] Sometimes also referred to as “associated registers”.

- 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

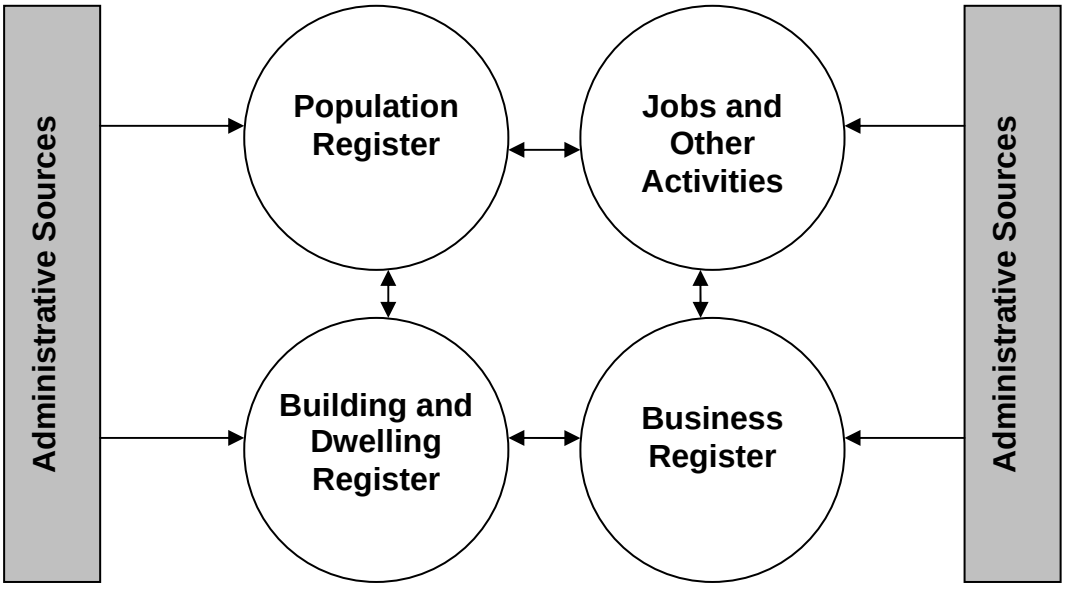
Examples of variables specific to the sub-set of units included in a satellite register could include “category” or “number of beds” for hotels, or “sales space” for retail businesses.

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.

8. 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.

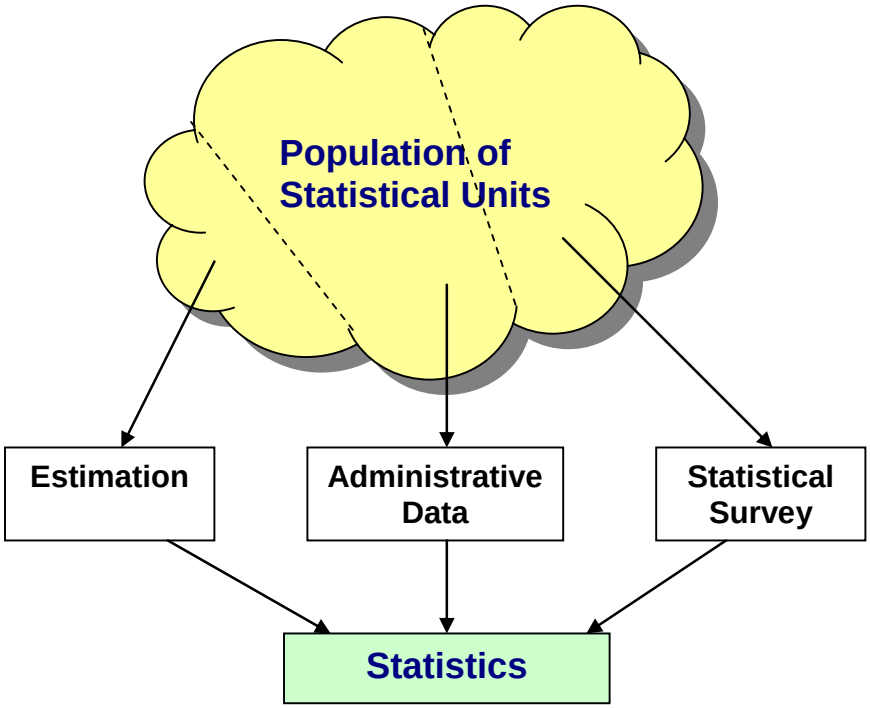
A practical example of the split population approach in business surveys is the Unified Enterprise Survey conducted by Statistics Canada. This brings together annual business data requirements, combining several previous surveys.

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%.[41]

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



[41] 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) 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
Summer Cottage Data	Q	Q/R	Q/R	R	R

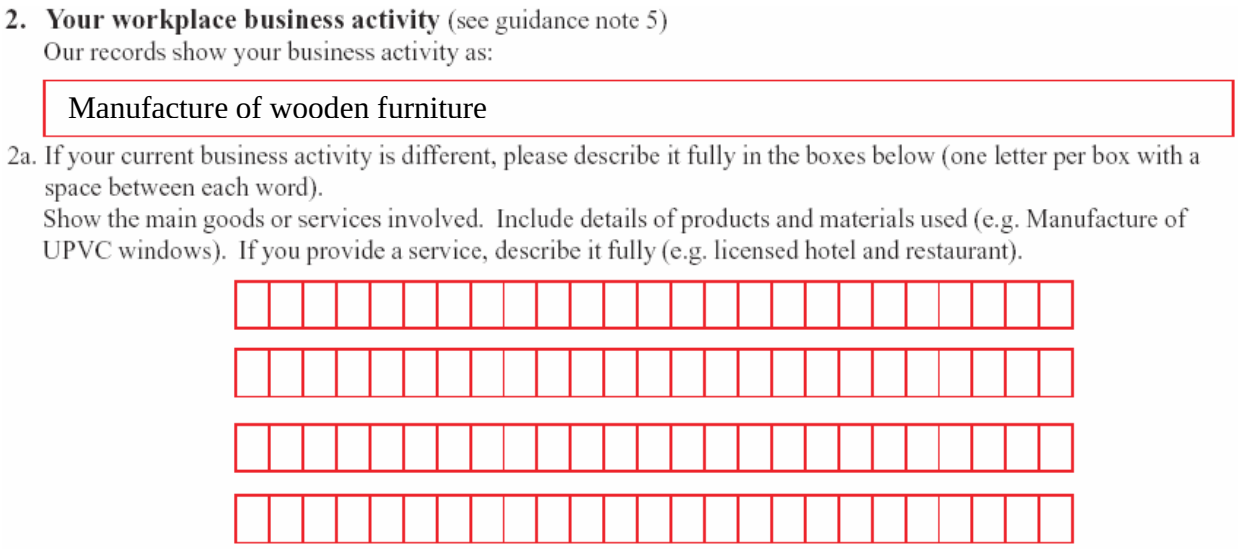
Key: Q = Statistical questionnaire
Q/R = Statistical questionnaire supplemented by administrative register
R/Q = Administrative register supplemented by statistical questionnaire
R = Administrative register
n/c = Not collected

Source: This table is a condensed version of Appendix 2 of the paper “Use of Registers and Administrative Data Sources for Statistical Purposes – Best Practices of Statistics Finland”: http://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10169.aspx

3) Pre-filled Questionnaires

This approach is really a special case of the split data model in that statistical questionnaires are still used to collect data about statistical units, but those questionnaires are pre-filled using administrative data as far as possible, with respondents being asked to merely check and correct these data where necessary. Figure 8.2 shows an excerpt from a pre-filled questionnaire, this example is from a British statistical survey of enterprises.

Figure 8.2 – An Excerpt from a Pre-filled Survey


2. **Your workplace business activity** (see guidance note 5)
Our records show your business activity as:

Manufacture of wooden furniture

2a. If your current business activity is different, please describe it fully in the boxes below (one letter per box with a space between each word).
Show the main goods or services involved. Include details of products and materials used (e.g. Manufacture of UPVC windows). If you provide a service, describe it fully (e.g. licensed hotel and restaurant).

Pre-filled questionnaires have three main benefits:

- They help reduce the burden on data suppliers by saving them time (assuming that checking and correcting is quicker than finding and entering data)
- They allow a check of the quality of the administrative data
- They are very flexible, particularly where the coverage of administrative data is not as complete as would be required for the other models

The main disadvantage is the risk of bias introduced because some respondents may simply accept the pre-filled data without checking them, or may choose not to spend time correcting errors.

4) Using Administrative Data for Non-responders

This approach can be seen as a variant of both the split source and the split data models. In this case, the statistical survey remains the primary means of data collection. However, statistical surveys tend to suffer from varying degrees of non-response, which affects the efficiency of the sampling process, and the quality of the resulting statistics. Non-response typically takes one of two forms, “unit non-

response", in which no data are supplied for the unit concerned, or "item non-response", in which a partial return is provided, but some data items are blank.

Dealing with non-response can be very costly for a statistical agency, as this typically involves repeat contacts by post or telephone to try to collect the missing data. This process is usually known as "response chasing", and tends to be very resource intensive.

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[42].

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[43]. In practice many examples of this approach concern using administrative data to improve estimates for small areas[44].

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.

[42] 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
[43] 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
[44] For example see: Using Administrative Records for Small Area Estimation in the American Community Survey, http://www.fcsm.gov/99papers/mcf.html

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.

9. Towards a Register-based Statistical System

9.1 Introduction

As stated at the end of Chapter 8, if administrative data are used to develop and maintain statistical registers, and also to supplement statistical surveys, the next logical step is to consider how to link those registers and surveys, and thus move towards a register-based statistical system. This approach has been developed mainly by statistical agencies in the Nordic countries, often with the initial focus of implementing a register-based population census.

A pure register-based statistical system could be defined as one in which all statistics (for a particular domain or set of domains) are produced exclusively from administrative sources that have been combined into two or more linked statistical registers. In practice, such a purely register-based statistical system is relatively rare, as small-scale statistical surveys are often needed for quality assessment or to overcome coverage issues for specific variables or sectors of the population. A more pragmatic approach is therefore to use the term "register-based statistical system" to refer to a system based primarily on administrative data that have been organized into linked statistical registers.

This chapter takes a brief look at some of the issues involved in the transition to a register-based statistical system. It intends to complement rather than duplicate the much more detailed study of this topic contained in the United Nations Economic Commission for Europe publication "Register-based Statistics in the Nordic Countries"[45]. That publication reviews best practices, with a focus on population and social statistics, and was prepared by experts from several Nordic counties, so should be regarded as the authoritative work on this topic.

9.2 Feasibility

Register-based statistical systems are not feasible for all countries, or even all domains of statistics, at least in the short-term. This is because the feasibility of developing and implementing such a system depends on a number of pre-conditions relating to policy and infrastructure, some of which have been mentioned in different contexts in preceding chapters. The key pre-conditions for a successful register-based statistical system are:

- The existence of suitable administrative sources: Comprehensive administrative registers of target populations are essential. The existence of large numbers of unregistered units, e.g. illegal immigrants or businesses operating in the informal economy, will make it extremely difficult to produce meaningful register-based statistics.

[45] See: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

- Ease of access – The administrative sources must be readily available to statisticians under the various frameworks described in Chapter 3. This includes the requirement that they are held in a format that facilitates data transfer.

- Common identifiers – Whilst Chapter 6 shows that common identification numbers for units that appear in multiple sources are not absolutely essential, they significantly facilitate the combining of those sources and therefore greatly increase the efficiency of production of register-based statistics.

- Public acceptance – As discussed in Chapter 4.2, the attitude of the general public to data linking and sharing within the government sector is a key factor in determining the extent to which administrative data can be used for statistical purposes. The balance between the efficiency of data sharing versus concerns about the protection of data relating to individual units is often the cause of fierce debate, with different outcomes depending on national cultures and traditions. In some countries the concept of a register-based statistical system is currently seen as unacceptable to large sections of the population.

If these pre-conditions are not in place, it is clearly not feasible to consider a register-based statistical system as a short-term option. This model could still, however be useful as a long-term goal, to be reached by following a step-by-step programme of changes to establish the necessary pre-conditions. Experience in the Nordic countries underlines the importance of long-term planning, as the implementation of register-based population censuses in those countries has typically taken around twenty years.

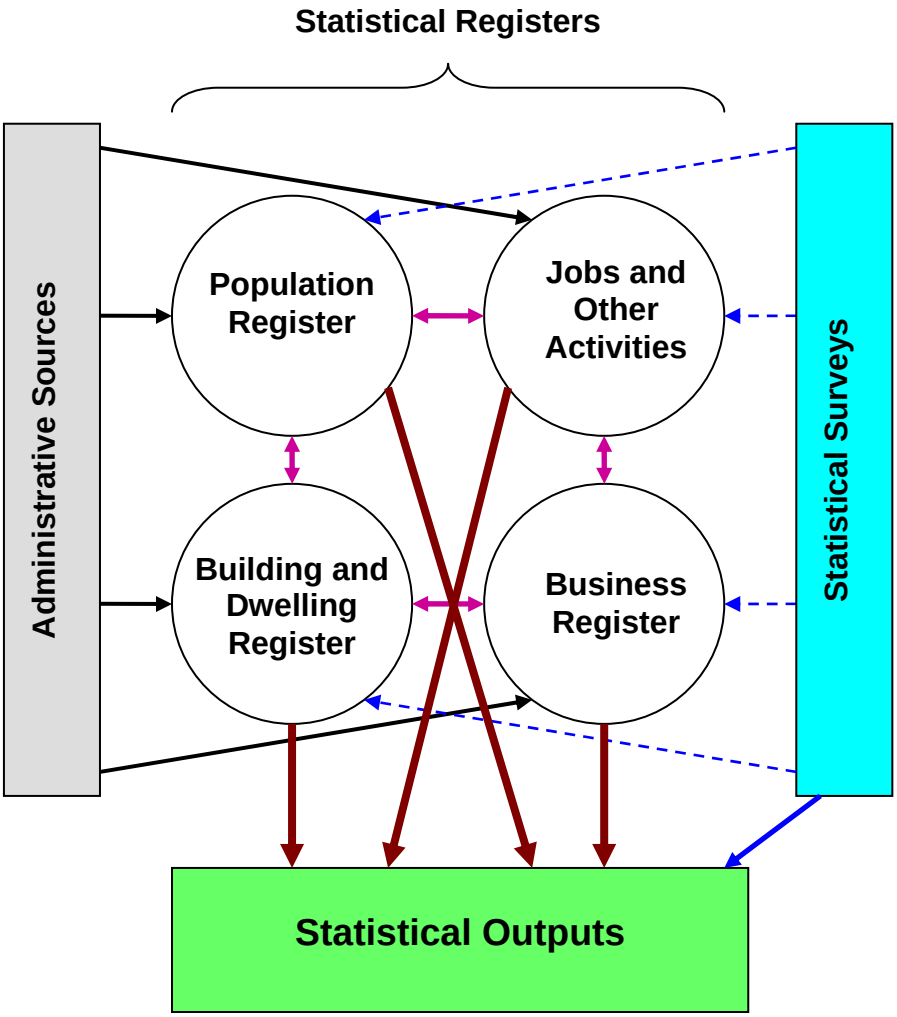
9.3 The Generic Model

Chapter 7.3 included discussion about register-based statistical systems insofar as they provide a model for the use of administrative data in statistical registers. Figure 7.4 in that chapter showed a generic model for a register-based statistical system, but focused only on the administrative inputs. Figure 9.1 below adapts that model to include statistical inputs and outputs. The two key features are:

- The links between the basic statistical registers - There may also be other, more specialized statistical registers, but these are not shown in the diagram for reasons of clarity.

- The balance between administrative sources and survey data in the statistical outputs - There is no clear rule determining what this balance should be, but it would be reasonable to expect that administrative sources are the main input.

Figure 9.1 – Register-based Statistical Systems – A Generic Model



Note: The statistical register of jobs and other activities is not always present in national versions of this model.

9.4 Summary

A register-based statistical system is clearly the ultimate goal when considering the greater use of administrative data for statistical purposes. In many countries it may seem a very distant goal, perhaps not attainable for many years. However, by adopting a strategic plan based on step-by-step improvements towards creating the necessary pre-conditions, it is possible to gradually move closer to this goal.