Using Administrative Data for Statistical Purposes

Stephen Penneck
Office for National Statistics, United Kingdom.

Introduction

This paper sets out some thoughts on the use of administrative data for statistics purposes from the point of view of a statistics producer. It is essentially the perspective of a central statistics agency, accepting that there are many uses of administrative systems for statistical purposes by administrative and policy departments not fully reflected here. Recent UK legislation gives us wider opportunities to access tax data to substitute for survey data.

In deciding how to pursue this advantage, some big questions arise:

- how will we manage data quality issues?
- what will it cost? will we save money?
- how will we manage perceptions around confidentiality?

In all these issues, Canada and Finland have vast experience which will help. The paper has four main sections. The first gives a general overview of the main types of sources for statistics. Further sections set out the UK experience with administrative data, current plans and some of the challenges.

1. Overview of Sources for Statistics

There are three primary sources of UK statistics: administrative systems, censuses and surveys.

Administrative statistics are the by-products of (usually) large scale administrative systems. By their very nature these statistics are often primarily used to manage those systems or measure their effectiveness and efficiency. Administrative statistics were among the first statistics produced by government, for example in the UK as compulsory registration of vital events took over from parish registers in the nineteenth century, and as trade statistics were generated by tariff administration and other trade controls. In a similar way all government departments and agencies collect information about their staff, expenditure, receipts, activity, output etc. While the predominant use is for management and monitoring purposes, many are turned into statistical series and published to enable the rest of government and the population at large to judge government performance and to contribute to a picture of the economy and society. ONS currently uses a range of these data from across government - e.g. NHS local registers for migration of patients between areas, and a sample of national insurance numbers for the Annual Survey of Hours and Earnings. In general these administrative systems are managed by government departments, and ONS requires a legal gateway to enable it to receive such data where it is disclosive (eg unit record data). These legal gateways do not generally allow ONS to receive substantive information from these records, just the basic information about the entity of interest. Thus most statistical analysis from these systems is produced by the departments themselves and this forms part of the wider UK statistical system.

Like most administrative sources, censuses aim to cover the whole population of interest. Thus the population censuses cover all people living in every residential address in the UK. The major advantage over administrative sources is that a census can be specifically designed to measure the population of concern, and each element in its design can be statistically controlled. Even a census is not fully comprehensive and needs to be supported by a coverage survey to provide a complete measure of the population. The major drawback of a census is the cost.

Sample surveys have developed as a whole branch of statistics in their own right. Statistical methods determine how representative samples are taken from a population of people, businesses etc and how attributes for the whole population are estimated, within known error or confidence limits. These methods are now used extensively by National Statistics Institutes and many other bodies worldwide. Not only do they enable the production of estimates at a known level of accuracy at a lower cost and more speedily than a census, but they can research an issue in greater depth. Interviewer expertise can be developed (for face to face and telephone collection) - a smaller number of observations collected well will have higher quality than a larger number collected poorly. Sample surveys are generally more flexible than administrative sources as they can be designed to meet a precise purpose as opposed to being the by-product of another system. However, they cannot produce precise detailed statistics and are highly dependent on high quality registers.

These three primary data sources are used to provide single source statistics in their own right - ie census reports and survey publications, but are also often combined to provide more complex secondary statistics - index numbers, labour market analysis and using conventional frameworks such as the national accounts and mid-year population estimates. Repeated in a consistent way over time these sources provide valuable time series, and most importantly statistical registers, which form the basis of sampling frames for surveys as well as a source for further analysis. Taken together they provide the source material for all our statistical analysis.

2. Current UK experience with administrative data

There are four main statistical uses of administrative data in the UK: analysis of administrative data as a single source, use linked with other sources for analysis, neighbourhood statistics, and in register building. Each of these is considered below:

2.1 Statistics which are a single source product of an administrative system are widely used in the management of that system and to assess policy options for change. Targets have become an important part of the assessment of government performance, and many of these targets - whether they are examination performance, hospital waiting lists or crime statistics - are statistics produced from the systems which administer these policies. Such statistics are often held in departmental management information systems and are used for internal management purposes as well as for providing measures for external accountability. This dual purpose leads to challenges of integrity, and in the UK both the Royal Statistical Society and the Statistics Commission have in recent years published reports on the use and challenges of statistics for performance indicators.

2.2 Two well known linked sources in the UK are the ONS and the Department of Work and Pensions' Longitudinal Studies. Both are good examples of how linked data can assist policy analysis.

The ONS Longitudinal Study (LS) contains linked census and vital event data for one per cent of the population of England and Wales. Information from the 1971, 1981, 1991 and 2001 Censuses has been linked, along with information on births, deaths and cancer registrations. At each census, data on slightly more than 500,000 sample members are added. During the 30 years of the study, around 1 million people have been recorded in the sample at some point.

The study was set up in the 1970s to meet the need for better data on mortality and fertility. Since then it has been used to address a wide range of research questions including studies of social mobility, ageing and migration. Studies that make the fullest use of LS data are those that link social, occupational and demographic information at successive censuses to data on vital events, such as studies of mortality, cancer incidence and survival, and fertility patterns.

Introduced in January 2004, and enhanced in October 2005, the Work and Pensions Longitudinal Study (WPLS) links benefit and programme information held by DWP on its customers, with employment records from Her Majesty's Revenue & Customs (HMRC). New data-sharing provisions introduced in the Employment Act 2002 enabled DWP to receive further data on employment from HMRC and use the information for wider purposes. DWP and HMRC have been working together to progress this initiative and to develop safeguards.

The WPLS offers DWP the opportunity to significantly improve both its analytical evidence base and its operational effectiveness. It supports the Department's agenda for child poverty, welfare-to-work and retirement income planning policy, and enables it to find out more about what works and what does not. This allows the department increasingly to target resources to the appropriate people, in the appropriate way.

2.3 The Neighbourhood Statistics website http://neighbourhood.statistics.gov.uk/ provides a powerful platform through which a wide range of high quality small area statistics is disseminated to an expanding audience of users involved in regeneration to the local public sector and to the wider population. Neighbourhood Statistics was developed following the Social Exclusion Unit's 1998 report on deprived neighbourhoods. That report recognised that government plans for the regeneration of the inner cities would be hampered by poor data availability. At the time there were few statistics available at a low geographic level. What did exist was held by individual departments and was not underpinned by consistent definitions or approach. Following a six year development programme, the small area statistical landscape has been transformed.

Administrative sources have been the backbone of the neighbourhood statistics development, enabling the website to provide analysis at local area level for: access to services; community well being; crime and safety; economic deprivation; education, skills and training; health and care; housing; population and

migration; physical environment; and work deprivation as well as the results of the 2001 population census.

2.4 Registers
Registers, requiring comprehensive coverage, are normally derived primarily from administrative systems but are often augmented by information from survey sources. ONS uses two basic registers whose main purpose is as a sampling frame for our various surveys.

a) The postcode address file. This is created by Royal Mail to plan postal delivery work. In the absence of a comprehensive population register, it is the most frequently used sampling frame for household and person surveys, although it depends on postal delivery staff to keep it up to date. It is a key asset as it allows stratification of samples by geography and clear identification of the address to target. It has little information about the other characteristics of addresses, and no information on who is resident at each address.

b) The Inter-Departmental Business Register. The IDBR is the comprehensive list of UK businesses that is used by government for statistical purposes. It provides a sampling frame for surveys of businesses carried out by the ONS and by other government departments. It has enough information, eg about size and industry, to enable efficient sampling stratified by these characteristics. It is also used in its own right to produce basic information about the structure of business in the UK and how it has changed over time.

The business register is based on inputs from three administrative sources: traders registered for Value Added Tax purposes with HM Revenue and Customs; employers operating a Pay As You Earn scheme; and incorporated businesses registered at Companies House. The ONS Business Register Survey and other surveys supplement these administrative sources, identifying and maintaining the business structures necessary to produce detailed industry and small area statistics.

The IDBR covers businesses in all parts of the economy, other than some very small businesses (self-employed, and those without employees and low turnover) and some non-profit making organisations. With 2.1 million businesses listed it provides nearly 99% coverage of UK economic activity.

3. Current plans and future directions

The Statistics and Registration Services Act 2007 provides a major opportunity to extend the availability of administrative sources for statistical purposes. It enables ONS to receive administrative data from other government departments, subject to the agreement of that department, provided Parliament agrees each specific case. The Act will come into force in April 2008, and the Government Statistical Service is beginning to develop a strategy for taking forward its provisions. Greater access to administrative data will bring significant benefits to UK statistics in a number of areas.
The main benefits are:
3.1 improved analytical capability in areas such as measures of economic activity, the labour market, pensions analysis, income and wealth, population and migration statistics, and measures of ethnicity and diversity

3.2 improved local area analysis the strength of administrative data in covering whole populations enables local area analysis to be produced to a level of detail not permitted by sample surveys

3.3 reduced costs to business replacing survey data with administrative sources enables sample sizes of surveys to be cut, reducing the form filling cost to industry.

Looking to the future it is clear that analysis of administrative sources at unit record level, often using linked data sets, is a growing area of competence for government statisticians. Its wide coverage and detail enables specific analysis of policy issues which aggregate survey results cannot address. Policy issues are increasingly cross cutting as interest grows in causal relationships across the economy and society.

Questions such as:

- why do some small businesses grow and others do not?
- what causes innovation?
- what are the determinants of poverty?

can only be answered by linking administrative data sets with each other and with survey data. While administrative data provide the impacts of policy measures, censuses and surveys provide the important demographic and structural characteristics that are needed.

There are two areas where the need for better use of administrative data is well established and is a priority for ONS: one to reduce business survey compliance costs; the other to improve population statistics. Both these uses will require integration of administrative sources with survey data at unit record level.
Integration of unit record business data into business surveys (mainly for small businesses) to replace survey

collection will be a new venture for the ONS. There is considerable international experience, in the Nordic countries, and in Canada, so we can learn much from them. We will need to look carefully at data editing and modeling routines to ensure we make the best use of these data.

ONS has conducted a limited feasibility study to look at using corporation tax records to replace some collection from the Annual Business Inquiry (ABI). The ABI is the main annual survey collecting structural data for the national accounts. It collects the main components of value added. In this respect it has similarities with corporation tax returns made to HM Revenue and Customs. ONS has obtained authorisation for a limited matching exercise for these two data sources, under confidential conditions, and the first results are now becoming available. These indicate that for smaller companies (those with a simple structure) there is a straightforward match for 85% of records, and with further matching effort, this proportion raises to 99%. For nearly half the cases where the units matched, the turnover figures were within 5%. However for a quarter of businesses the differences exceeded 25%. Further work on the causes of the major differences is continuing. If this work is successful we can use tax data to reduce substantially the statistical reporting burden for smaller companies.

A second initiative is the new population statistics strategy which aims to provide a long term vision for population statistics. The decennial Census traditionally provides benchmark population statistics updated with mid year estimates from registration sources. But this has not been robust in the face of population changes and changing user needs. Populations have become more mobile and residency arrangements and household structures have become more complex. In addition there are needs for increased frequency, and a more flexible counting base (usual residence; daytime/service population; etc). Users require improved accuracy and more confidence in the estimates and have a strong demand for more small area statistics. The main components of the proposed system are:

- A high quality address register
- A possible population register
- The integrated household survey integrating continuous household survey data
- A linked statistical database, linking administrative and survey data at individual and household level
- A full Census for 2011 which potentially enables census and statistical databases to be linked to create a population statistics database

4. The statistical challenges

Given the essential differences between administrative statistics and those based on survey samples, there are inevitably some importance differences in their attributes, their quality and their appropriateness for particular purposes. There are important challenges in how a statistics office manages data quality issues; whether this will add significantly to cost or save money; and issues of public perception of confidentiality.

The most important defining aspect of administrative statistics is their linkage with that administrative system. That linkage brings advantages and drawbacks. While the resulting statistics will be highly relevant to the management of that system, attempts to use them for broader purposes immediately lead to quality issues which can be expensive to remedy, and can lead to misleading conclusions. A major perceived advantage of administrative data is that they might be essentially cost free. The administrative system exists for its own purpose and bears its own cost. Producing summary statistics often requires minimal marginal cost. This saving can be illusory, especially if the statistics are used to measure aspects not closely related to the domain covered by the administrative system. So a social security system can be used to produce statistics showing the operation of that system, at low cost, but attempts to use those statistics to derive measures of poverty will usually require additional cleaning, and possibly matching with other data sources, or modeling, which can all be expensive and be testing for the quality of those estimates.

The second major advantage is that administrative statistics usually cover the whole of their particular population. This allows analysis of small population groups and is especially useful for those interested in rare populations, small geographies and local area information. However, the population measured is that covered by the administrative system, which may not be the population of interest for analytical purposes. Although administrative databases can be large and unwieldy for analysis, this can be overcome by sampling.

Administrative statistics will be regularly updated by the administrative system, but their timeliness will be driven by the needs of that system. Tax records can be slow; birth registrations can be quick. Essentially the close linkage with the administrative system, while

potentially bringing benefits, can also be a significant drawback. All the attributes of administrative statistics are set by the needs of the administrative system. As a by-product, analytical needs have very little influence over any aspect of the system. Thus questions of definition, units, classification, coverage, methodology, frequency and timing are determined necessarily by consideration of the administrative system of interest. Changes in policy and administrative practice can have serious implications for resulting statistical measures and time series. Adding further items of information, improving data cleaning or changing definitions will often improve statistical quality, but will usually impose large administrative costs which the owner of the administrative system is often unwilling to bear. If systems fail (as has recently happened with birth and death registrations in the UK) the departmental priority will be to fix the administrative system leaving statistical uses until later. It is more important to register deaths quickly than to produce the statistical analysis of deaths.

By contrast sample surveys are usually specifically designed for analytical purpose, so the coverage, definitions, methodology and timing can be designed to meet analytical needs. However, sample sizes can be small - large scale surveys are expensive, and small scale surveys have limited use for analysing small populations or local areas. Samples are subject to sampling error and non-response bias. Non-response bias is partly related to response rates, and household survey response rates in the UK have been falling over the last ten years, raising concerns about the continued accuracy of survey outputs. In addition, we cannot be certain of the accuracy of business survey responses, compared for example, with administrative data collected for tax purposes. Furthermore, surveys impose some compliance cost on respondents - whether they are statutory surveys of businesses or voluntary surveys of individuals. Administrative systems may include some collection of data from individuals (eg medical records) but the individuals concerned will see this as a necessary part of the administrative process rather an additional statistical burden.

So a key issue is how much influence statisticians can have on the design and operation of administrative systems. In international discussions statisticians from other countries are often negative about the extent of this influence.

We are in a better position in the UK. UK Statisticians work in policy departments on administrative data sets. They are in a good position to influence the characteristics of those data sets. And they are in the Government Statistical Service - so can appreciate wider statistical needs across government.

The advantages and disadvantages of these two sources can be summarised according to the different dimensions of quality as follows:

Table: Illustrating some of the different aspects of quality for administrative and survey sources

Dimensions of quality	Administrative data	Sample surveys
Relevance	Definitions and coverage will be relevant to the administrative system, rather than the analytical need. Good source for detailed and local area analysis.	Surveys can be designed to be relevant to the analytical need. Quality for detailed analysis is constrained by sample sizes
Accuracy	Subject to non-sampling error. Not under the control of statisticians	Subject to sampling as well as non-sampling error. Under statistical control
Timeliness	Some sources (eg tax data) less timely than surveys	Many administrative sources very quick. Surveys subject to response times.
Accessibility	Depends on legal structure. May also be technical and institutional barriers	Under direct control of the statistical agency
Comparability	Dependent on changing administrative definitions over time	Under direct control of the statistical agency
Coherence	Often enables data linking if common identifiers exist	Depends on common registers

Statistical quality measurement is based on a set of well understood techniques. Non-sampling errors are often measured through analysis of the process or

external comparisons with other data sources. These measures are not always available for administrative data sources, making the measurement of their quality more problematic. In addition little may be known about the quality of new potential statistical sources, and how difficult this will be to assess.

It is not surprising that many statistical agencies have often favoured the use of surveys over administrative data given the greater control possible over quality, and the difficulties some have in gaining access to administrative records. But at a time when governments are particularly looking at reducing the cost that they impose on society and business, there is a continued drive to reduce survey compliance costs. The UK has a good record of measuring and reducing these costs, but is looking for a significant further reduction of 25% over the next ten years; a reduction which can only be achieved by replacing some survey data with administrative data if adequate statistical outputs are to be maintained. The challenge for the statistician is how to achieve this without losing statistical control over definition and methods and without lowering statistical quality.

A further challenge concerns perceptions of confidentiality. Many of the strongest benefits for statistical analysis come when it is possible to link administrative records from different systems together, or when they can be linked with population censuses or surveys. All this linking must be done under statistically controlled conditions, but it offers powerful analysis of some of the cross cutting policy issues of wide interest to governments and the public. Linking censuses and surveys with tax and benefit records provides an analysis of the demographic and household composition for those in poverty. Linking with educational achievement records provides insights into one cause of poverty. These linking studies require a common identifier to link the records, which is not always available for the relevant data sets. It also requires a public understanding of the way that these records are being linked and the differences between linking data sets for statistical purposes and linking for administrative or policy purposes. When we link for statistical purposes, all linking takes place within a statistical authority and the only information to leave that protected domain is in the form of non-disclosive analysis, providing useful insights into characteristics of the population or businesses, but not providing any information which could lead to any disclosive inferences being drawn about individuals or businesses. This is not well understood by the public, and with the UK media active on these issues, statisticians need to proceed with caution.

5. Conclusions

It is clear that administrative sources, censuses and surveys each have their own strengths and weaknesses. My impression is that the UK statistical system is more survey reliant than many other countries and has traditionally made less use of administrative data for wider statistical purposes, partly because of the lack of a comprehensive population register. With a growing focus on reducing compliance costs; increasing interest in local area information; and in cross cutting policy analysis requiring data linkage, ONS is reassessing the place of administrative data in the UK statistical system. The new Statistics and Registration Services Act provides the means for this.

However we need to continue to bear in mind the relative quality attributes of administrative and survey data. Increased use of administrative data will require quality issues to be addressed and this may mean that the idea of administrative data as a 'cost free' source may be illusory. Solutions will need to be sought in two ways. Firstly, it will be important for statisticians to work more closely and have more influence with the owners of administrative sources. This is the only way to ensure the maximum analytical use of these sources, through the use of common statistical definitions and classifications. In the UK, Government statisticians are well placed to exert this influence. Secondly, statisticians will need to develop and invest more faith in automatic editing procedures which clean data without requiring contact with the data subject. This will be especially important for variables which are of little interest to the administrative data owner, but of greater interest to the analyst.

Finally is the issue of public perception of confidentiality. This may limit the speed with which we can progress, building trust as we go.

References

1999 Social Exclusion Unit: PAT 18 Better Access to information
2003 Royal Statistical Society: Performance Indicators, good, bad and ugly
2004 O'Donnell Review: Financing Britain's future - review of the Revenue Departments
2004 Allsopp: Review of Economic Statistics for Economic policy making
2005 Hampton: Reducing administrative burdens
2005 Cabinet Office: Transformational Government
2006 Statistics Commission: PSA targets - the devil in the detail
2006 ONS: Neighbourhood Statistics Service annual report to ministers 2005/6.