

Regional Training on Producing Register-based Population
Statistics in Developing Countries
23 September – 31 October 2013

e-learning module: *Basic information and statistical background*
23 – 27 September 2013

Lesson 1

Introduction to administrative sources; *basic concepts, uses and quality aspects*

References and complementary readings:

1. United Nations (2011): [Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices](#)
 2. United Nations (2007): [Register-based Statistics in the Nordic countries](#)
 3. Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*, John Wiley & Sons, Ltd, Chichester, UK.
-

Contents

1. What are administrative data?
 2. Why administrative data?
 3. Quality of administrative data
-

1. What are administrative data?

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 paper¹. 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:

“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. 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: (i) increasing privatisation of government functions, (ii)

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).

growth of private sector data and “value-added re-sellers”, and (iii) user interest in new types of data.

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.

- 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.

The definition proposed by Eurostat is therefore:

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

What is a register?

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.

However, in some broader definitions (Wallgren & Wallgren, 2007) censuses are considered to be source of data for statistical registers since they are comprehensive and make updating of unit-level information possible.

A register is defined in “Register-based Statistics in the Nordic countries” as *a systematic collection of unit-level data organized in such a way that updating is possible*. Updating is the processing of identifiable information with the purpose of establishing, bringing up to date, correcting or extending the register, i.e. keeping track of any changes in the data describing the units and their attributes.

As a rule, a register will contain information on a complete group of units, a target population (e.g. persons, buildings, firms). These units are defined by a precise set of rules (for instance resident population in a country), and the attributes are updated in line with changes undergone by the units.

The definition of registers given here refers exclusively to the content of the data collections we are dealing with. Information in registers can be stored in a variety of ways.

A key requirement is that each unit in the register can always be uniquely identified. This is best achieved by using a system of *identification codes* (keys), but identification is also possible without such a code if sufficient information on the units is available (for persons: name, address, date of birth etc).

Ideally, for statistical use, registers should be as *comprehensive* and *accurate* as possible, but even incomplete registers can be used for statistical purposes.

Basically a register may be a list of all units in the population and nothing more. In practice most registers also comprise some additional attributes for each unit (data). The term data register is used to separate registers with data from the mere lists of units. Here, the term register is used in both senses.

“*Administrative registers*” are registers primarily used in administrative information systems. That means that the registers are used in the production of goods and services in public or private institutions or companies, or that the information is a result of such production.

The term “*Statistical register*”, in its broad definition, describes a register created by processing data from either census conducted by the statistical office or on administrative registers from authorities and organisations outside the statistical office.

2. Why administrative data?

According to the principle 5 of the Fundamental Principles of Official Statistics, adopted by the Statistical Commission in 1994:

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.

Most national statistical offices (NSOs) seem to have experienced budget cuts, without receiving any priorities or signals on how to reduce the statistical programme or maintain the quality of the statistics. Instead we have observed a growing interest in statistics on new subjects and on new combinations of variables. In many countries there has been a growing demand for statistics on small population groups more frequent than the ten-year based population censuses. At the same time, NSOs have duty, according to the fundamental principles of official statistics, to produce official statistics with highest possible quality.

Doing more with less resources will only be possible with shifting production paradigm from traditional methods of data collection to efficient use of already available data records that can be used for statistical purposes.

1.1 Use of administrative data in production of statistics

There are 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.

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 followed-up, responses have to be processed and verified, and results have to be calculated. Traditional censuses are even worse because they are conducted on a much larger scale.

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.

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.

Thus, if policy makers and respondents are united in calling for reductions in the statistical response burden, it is extremely difficult for NSOs to resist this pressure, and the re-use of data collected by others is the logical solution.

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. 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.

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.

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.. 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.

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 NSO 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.

3. Quality of administrative data

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 list²:

- *Relevance* - the degree to which statistics meet the needs of current and potential users.
- *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.

² 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.

-
- *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, comparability through space, and comparability between domains.

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.

Statisticians always have to make compromises with their ideals in order to get practical results from the data collection. In a way, the use of administrative data has rearranged the situation: We have the product of a data collection process, and have to compare that product with our quality requirements for the statistics to see if the difference is acceptable. The administrative definitions of the target population may not correspond to our needs (employees, but not self-employed, in employment registers), the variables are not always defined in the way we want (de jure instead of de facto place of residence in the population register), and the time references are not always as precise as we want them to be (the time references of the register may not coincide with the "census day" of the statistics). However as these data are almost free for use for the NSOs, we can use our resources to supplement the information not covered by the administrative data, and make corrections just as we have to do when we are using traditional methods.

When planning a census or a sample survey, the NSOs have some ideas of quality when defining units and variables. In practice, and despite all efforts that we make, sampling and non-sampling errors remain. We must at all times strike a balance between the quality wanted and the practical and economical realities.

Administrative data are produced on the basis of some administrative processes, and units and variables are defined out of administrative rules and demands. The definitions may differ

from the needs of the official statistics, but the data are usually of good quality for their administrative purposes.

In simple terms: Using traditional data collection methods, the NSO is asking: "*We asked our question, but do we get the right answer?*" Using register-based (administrative) data the NSO is asking: "*Surely we got the correct answer, but how does it correspond to our question?*" The NSO is working towards the same goal but from a different starting point.

However, it is important to note that administrative data records can never fully replace direct data collection by sample surveys, but that these two methods complement each other.

Summing up the result may be that for a given amount of money the quality requirements are better fulfilled by using data from administrative registers than by traditional data collection by means of questionnaires.