Using administrative data in population and social statistics

Jari Tarkoma

Statistics Finland, Helsinki, Finland jari.tarkoma@stat.fi

Abstract

The use of administrative data in statistics production is increasing everywhere. In many countries it is common practice especially within population and social statistics. There are several well-known reasons and benefits for the production of register-based official statistics. However, there are still concerns regarding the quality of statistics based on administrative data. This presentation will briefly describe the preconditions, challenges, reasons and possibilities for register-based statistics production. There are questions concerning availability, standardisation of concepts and definitions, limitations, quality of data and comparability of statistics. For instance, there are no similar methodological framework and similar standard quality criteria and quality indicators for register-based statistics as for survey statistics. In order to justify the use of administrative data for statistical purposes, statistical authorities have to face legal, contentual, technical and methodological issues. There are some important issues to focus on, e.g. cooperation among the administrative authorities and possible coordination of standardisation activities concerning administrative records, quality assurance and reliability studies of basic administrative data and comparison of surveys and register-based statistics. If these aspects are properly taken into account, there should be no barriers for using administrative data and gaining advantages of this.

Key words: register-based statistics, official statistics, quality

1. Introduction

In Finland, the exploitation of data from administrative sources in statistics production began in the 1970 Population and Housing Census. Population registers were established in the Nordic countries in the 1960's and the personal identification number was introduced at that time too. The register-based statistics production first started in population and social statistics, but nowadays administrative sources are also widely exploited in business and environmental statistics. Statistics Finland has produced census data annually since 1990. All demographic statistics are compiled directly from the Population Register (formally Population Information System).

The use of administrative data in statistics production is increasing all the time. It is common practice especially within population and social statistics. There are important reasons behind this trend: the increased demand of high quality statistics, and at the same time the pressure to reduce both the costs of data collection and the response burden for citizens. The situation varies in different countries depending on the historical development of registers and administrative data sources. The possibilities are always dependent on the national circumstances and the current challenges faced in the countries may also differ. In the last Population and Housing Census round in 2010 and 2011 at least 16 countries of today's European Union used registers as one of the data sources. It is essential to note that the use of administrative data will usually increase gradually, step by step. The progress may be slow, if there is a lack of basic preconditions. For instance, the possibility to link data from different sources using the personal identification code is a crucial part of all statistics production.

Administrative data are collected for administrative purposes and they are linked to different kinds of administrative processes and decision making in the society (population registration, taxation, social benefits, administrative permits, pensions). The main idea behind using administrative data is that these data already exist for administrative purposes and it is possible to use them for statistical purposes. What is the best way to utilise these potential data is not always a simple question and must be decided case by case in the modern complex society. As regards the use of administrative data for statistical purposes, it is inevitable that, e.g. the cost-benefit analysis has an essential role. On the other hand, statistical authorities have to encounter many questions concerning the quality of the data and justify the use of administrative data for statistical purposes. The definitions used by administrative authorities may differ from the needs of statistical offices, but the data are usually of high quality for administrative purposes. However, often the non-statistical and statistical definitions are close enough and the similarity is very high.

2. Some preconditions, reasons and ways of using administrative data

There are several well-known reasons and benefits for the production of register-based official statistics. Sample surveys are often an expensive way of data collection. At the moment, all national statistical offices are trying to lower the costs of data collection and reduce the response burden. They all also struggle with the low response rates in social surveys. Using administrative data usually means a possibility to produce statistics annually instead of with a five or ten year interval. Moreover, dealing with totals (complete or almost complete coverage) instead of samples often makes it possible to produce longitudinal statistics and high quality small-area statistics or even grid-based statistics. The use of the coordinate-based statistics (geocoding of all relevant units in registers used for statistics and integrating GIS with statistical data) and modern technical solutions has already opened new effective ways of producing, utilising, disseminating and visualising statistics.

Among the most important general preconditions for register-based statistics production are the legal basis and public approval in the society. A legislation that enables and supports the use of registers is needed. The Finnish Statistics Act takes into account both the cost-benefit analysis of data collection and the response burden. It guarantees the access to administrative data for statistical authorities. Government agencies have an obligation to deliver data to Statistics Finland. The Statistics Act (2004, originally 1994) is based on the principle and obligation that whenever possible, statistics shall be compiled using existing administrative records and not to re-collect the data. Before starting a new direct data collection, statistical authorities are obliged to examine whether these data already exist in administrative sources. The Statistics Act authorises Statistics Finland to access administrative data on unit level with identification data and to link data from different sources for statistical purposes. As regards the public approval to use register-based data in statistics production, it is extremely important that the obligation of data protection and confidentiality are included in the Statistics Act and that the procedures concerning these matters are transparent, sufficient and effective. There is also a special Personal Data Act (533/1999) concerning the use of personal data.

If there are high quality administrative data available in the society, there are many different ways to utilise these data. Firstly, the most straightforward way is direct use, which simply means to obtain the data directly from the register (age, gender, marital status, citizenship, income data, etc). Secondly, a more complex way is the so-called register-estimation in order to form completely new variables using data in several registers. The aim is to estimate for each statistical unit the value of the target variable as close to the statistical concept and definition as possible. This is done by using all existing relevant data available and a set of decision rules to estimate the value of the

statistical variable (e.g. the deduction of the main type of activity of a person using about 30 different administrative registers. Thirdly, the combined use of survey and register data is very common. This includes different methods: to use some additional administrative data in a survey, to use register data as a sampling frame in a survey, to use administrative data in non-response control, to supply variables for units that do not respond to a statistical survey, to evaluate the structure of non-response (non-response analysis), and to use administrative data in imputation. Utilising administrative data also makes it possible to improve the accuracy of statistical estimation.

3. Some challenges and quality issues

In spite of the wide exploitation of administrative data, there are not similar methodological framework and similar standard quality criteria for register-based statistics as for survey statistics. To produce sample based social statistics is quite different from producing statistics based on administrative data and total populations. However, there is also a need for methodological framework in register-based statistics for dealing with quality and combining various administrative sources and surveys (linking, editing, imputing, estimating, modelling, deriving of new variables, etc.). Therefore, quality and usability of administrative data should always be analysed carefully. This detailed analysis is partly based on the recommendations and official classifications used in population and social statistics.

There are some disadvantages concerning the use of administrative data, which have to be considered carefully. Firstly, the concepts and definitions may differ from the statistical ones, there may be different reference periods, various dimensions of quality, changes in time, etc. There may also be a limited data content, because only variables covered by registers are available. Secondly, the dependence on data suppliers (administrative authorities) and possible changes in the data content or technical solutions of registers can be problematic. This means a vulnerability to changes in legislation and administrative practices. Thirdly, the coverage of registers may be defective for some data, some data may not be included in the register at all and there may be quality problems. For instance, consistency problems may arise when linking data from different sources.

In Finland the use of administrative data has been connected to the attempts to harmonise register-based statistics with survey-based statistics and to ensure that these are as comparable with each other as possible. The use of administrative data relies quite naturally on output-harmonisation. In this work, it is important to verify the differences in the definitions of variables between administrative and statistical systems and to understand these differences and their impacts. In order to ensure the quality of register-based statistics, some extensive studies were carried out in connection with the 1985 and 1990 Censuses and the data were compared both at aggregated level and at unit level. Differences between the administrative data and statistical concepts are assessed at the level of individual observations and measured at aggregate level, whenever data is available and allows that.

The question about the quality of administrative data arises in three different points of the statistical process: 1) Quality of administrative data itself as a source data; 2) Quality of data processing; and 3) Quality of produced statistics. The second point has much to do with the editing and imputing processes and other advanced methods to process the data, e.g. data matching. The following chapters will shortly describe the cooperation among the administrative authorities, monitoring of the quality of Population register, and the possibilities to compare register-based data with survey data. These are all examples of continuous quality improvement and assurance.

4. Cooperation with administrative authorities

Close and well-functioning cooperation between the statistical and register authorities is essential for the effective use of administrative data and for the continuous quality assurance. This cooperation has a long tradition in Finland and it is a part of the coordination system of the Finnish official statistics. The use of administrative data requires cooperation on the highest level of authorities. There is a special cooperation forum for national register authorities, and regular bilateral negotiations for the change of information. The basis for the good cooperation between statistical and administrative authorities is the Finnish Statistics Act.

Coordination is centralised in Statistics Finland, there are regularly meetings on various levels of the different organisations. The commitments are made on the highest possible level. There are also named contact persons for each register authority and each register. The possibilities to influence the content of administrative data are persuasive. However, there is a need for continuous monitoring of the changes and the work is proactive by nature. The possible impacts of changes in legislation on statistics production are always discussed beforehand. During the last decades there has not been many changes in the registers with a negative influence on statistics production.

It is sometimes suggested that the statistical authorities should have a clearer role in the coordination of the registers and administrative data sources from a statistical point of view. The forthcoming updated Regulation on European Statistics (EC) No 223/2009 may include more possibilities for the statistical authorities as regards the standardisation activities concerning administrative records that are relevant for the statistics production.

5. Quality assurance and reliability studies of administrative data

The administrative authorities are responsible for the quality of their administrative data. The use of administrative data presupposes that the content and quality of these data are known. There is need for continuous assessment of the quality of administrative data before statistical use. It is both economic and reasonable to try to control and manage the quality issues at the source of administrative data. There are often several users of the administrative data and quality control will also diminish the need for editing and imputing the data (missing values, over/undercoverage, etc.). The register authority and statistical authority have a shared interest to guarantee the high quality of the register data.

One possibility to control the quality of registers is to carry out quality check surveys e.g. to use ad hoc questions in a survey. The Population Register is one of the main sources for population and social statistics. Population data are also used as a sample frame for many important social surveys and as complementary data in these surveys. Population Register Centre (the authority responsible for the Population Register) monitors the quality every year through a survey in order to establish the accuracy of address data recorded in the register. This data is the only source of addresses for persons in surveys and the statistics on families and households are also based on this data. In connection with the Labour Force Survey some data in the Population Register are checked once a year. This quality survey has been carried out since 1998, last time in November 2012. There was an additional question about the quality and accuracy of the address data in the Population Register. The proportion of people with a correct address in the register has always been high, over 98 percent of the respondents. Some other data, e.g. native language, occupation and tenure status of dwelling have also been checked in recent years.

A mixed use of administrative data sources will reveal differences between these sources. This kind of comparison may also provide important information on the quality of these sources and often leads to a selective use of administrative data.

6. Comparison of surveys and register-based statistics

Combining data from surveys and registers offers opportunities for interesting quality studies. Sample surveys may be used for quality assessment of register-based statistics and vice versa.

Statistics Finland uses the Labour Force Survey (LFS) data to evaluate the quality of the annual Register-based Employment Statistics (RES). The LFS is based on a random sample and the monthly sample consists of about 11,000 people aged between 15 and 74. LFS is utilised in two ways. Firstly, to monitore the level of the statistics. Secondly, the individual data of the LFS and the annual RES are matched on individual level by using the personal identification numbers. By using the compiled dataset, it is possible to make comparisons between these two sources and analyse the reasons for differences in them. Comparison of the RES and LFS describes the differences between these two statistics on individual level.

Table 1: Main type of activity according to the Register-based Employment Statistics (RES) and Labour Force Survey (LFS) in December 2010 in Finland (persons)

	LFS							
	Total	Employed	Unemployed	Students	Pensioners	Conscripts	Others	Non response
RES								
Total	**9,295**	5,346	432	915	2,051	42	509	2,848
Employed	5,303	**4,994**	45	76	50	1	137	1,474
Unemployed	556	83	**274**	8	17	-	174	252
Students	1,036	107	73	**811**	9	1	35	306
Pensioners	2,079	117	7	1	**1,941**	-	13	524
Conscripts	40	-	1	-	-	**39**	-	10
Others	281	45	32	19	34	1	**150**	282

Table 2: Main type of activity according to the Register-based Employment Statistics (RES) and Labour Force Survey (LFS) in December 2010 in Finland (per cent)

	LFS							
	Total	Employed	Unemployed	Students	Pensioners	Conscripts	Others	Non response
RES								
Total	**100.0**	44.0	3.6	7.5	16.9	0.3	4.2	23.5
Employed	100.0	**73.7**	0.7	1.1	0.7	-	2.0	21.8
Unemployed	100.0	10.3	**33.9**	1.0	2.1	-	21.5	31.2
Students	100.0	8.0	5.4	**60.4**	0.7	0.1	2.6	22.8
Pensioners	100.0	4.5	0.3	-	**74.6**	-	0.5	20.1
Conscripts	100.0	-	2.0	-	-	**78.0**	-	20.0
Others	100.0	8.0	5.7	3.4	6.0	0.2	**26.6**	50.1

There are many reasons for the observable differences. By using more detailed register-data, it is possible to analyse the reasons for most of them. For instance, unemployment in the RES is based on the official register of the Ministry of Employment and Economy, in the LFS respondents are defined according to their own report. The table reveals much of the structure of the non-response in the LFS. The proportion of persons classified in the same way in both statistics is 67.6 per cent. However, if non-response is not included, the proportion of persons classified in the same way is 88.3 per cent.

7. Conclusions and future

There is an increasing demand in many countries to try to cut down the costs of statistics production whenever possible. The best way to do this is to find innovative solutions and new data sources in order to produce high quality official statistics. Statistical authorities have to be proactive and ensure that the statistical point of view is sufficiently taken into account in registers and administrative data. There is no doubt that the use of administrative data will increase and the different ways of utilising it will become more of a standard in producing statistics everywhere. Also, the domain of administrative data will expand to the area of private sector electronic datasets and structural big data.

The use of new administrative data must be based on detailed analysis of these data. This means piloting and testing, clear and distinct consideration, comparing the statistical and administrative concepts and definitions, and making sure that the quality of statistics is high enough. Administrative data collection cannot always replace sample surveys as an instrument for data collection. All population and social statistics can never be based only on administrative data. These two methods should not be seen as rivals, but as complementing each other.

The efficient and broad use of administrative data for statistics production presupposes an integrated system of population and social statistics. In Statistics Finland, this means a data warehouse solution, common metadata, and an attempt to build an integrated system for the use of administrative and survey data with advanced editing and imputing methods and continuous quality assurance. Using administrative data in an advanced way in the production of population and social statistics makes it possible to compile high quality statistics cost-effectively, to combine sample surveys and administrative data, but also to develop statistics for the new and current needs of society.

References

* Heimonen, J. (1994). Evaluation Study of the 1990 Census. Statistics Finland, Population Census, Volume 9B. Helsinki 1994.
* Hokka P. (2012). Reliability Study of the Population Information System 2012 (unpublished paper only in Finnish).
* Korpi H. (1989). Main type of activity and occupational status in the 1985 census. Register-based parallel data. Statistics Finland, Studies nr. 152, Helsinki 1989.
* Use of Registers and Administrative Sources for Statistical Purposes, Best Practices of Statistics Finland. Handbooks 45, Helsinki 2004.
* Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics. United Nations Economic Commission for Europe (2007).
* Using Administrative and Secondary Sources for Official Statistics. A Handbook of Principles and Practices. United Nations Ecnomic Commission for Europe (2011).
* Wallgren, A., Wallgren, B. (2007). Register-based statistics: Administrative Data for Statistical Purposes. Wiley Series in Survey Methohdology, John Wiley & Sons, Ltd, Chichester, England.