

We must use administrative data for official statistics – but how should we use them?*

Eivind Hoffmann

Bureau of Statistics, International Labour Office, 4 Rue des Morillons, CH-1211 Geneva 22, Switzerland

This paper argues that there are both inherent and operational issues of quality associated with the use of administrative registrations as basis for official statistics, with respect to coverage, timeliness, frequency, validity, reliability and consistency. It further argues that because increased indirect and direct use of administrative registrations seems inevitable, the official statisticians must apply and adapt the established practices and principles of their trade to get to know the data generating process in detail, to monitor the data collection process, to try to persuade the responsible agencies to make changes which lead to improvements in data quality, to calibrate the observations generated by the administrative registrations by the use of statistical surveys, to use the administrative data and the calibration results as a basis for estimating the statistical parameters to be published, and to explain to the user how to properly interpret the resulting statistics. Procedural and inherent weaknesses of administrative registrations can only be overcome by developing and using appropriate methodological counter-measures.

1. Introduction

The advantages and problems involved in using administrative data as a basis for official statistics should be well known to everyone producing such statistics, see, e.g., [1,2]. The advantages, in particular from the perspective of the budgets of the statistical agencies, are in first instance related to what one may call the "golden rule"

*Based on a paper originally presented at the First Conference of the International Association of Official Statisticians, Rome, October 1988, and included in its *Proceedings*. The permission of the International Statistical Institute to make use of the paper in this form is gratefully acknowledged. Some of the modifications were inspired by the discussions at the *ECE/CEE Work Session on Registers and Administrative Records for Social and Demographic Statistics, Geneva, 23-25 January 1995*. The article can be seen as a first attempt to create a framework for the forthcoming ILO work to develop guidelines for the use of administrative registrations as basis for labour statistics. However, the views expressed are those of the author and do not necessarily reflect those of the ILO or its Bureau of Statistics.

of official statistics, namely that, *ceteris paribus*, the cost of data collection increases with the number of respondents who must be contacted. Therefore, if the statistical agency can get information from agencies which have already collected it for their own use, that will almost always appear to be less expensive than to collect corresponding information from the units of interest themselves. However, whenever the use of administrative records as a basis for statistics is discussed the advantages need to be seen in the light of the possible problems which this strategy may represent, in order to ensure that appropriate efforts can be made to identify, describe and counteract whatever deficiencies there may be in such data, and to explain the consequences of these deficiencies for the use of the statistics produced. Not to make these efforts damages the professional reputation and integrity of the official statistician. The objective of this article is to remind us about the issues involved and to present elements of a strategy to deal with them. The issues discussed are most urgent when the administrative registrations *substitute for* a statistical survey of the units of observation, but should also be carefully considered when using the administrative registrations to *supplement* a survey or to *support* the planning, processing or control of a survey.

2. Quality Issues

The quality issues related to the use of administrative records as basis for statistics arise from the simple fact that administrative data are collected as part of an administrative process, and the consequences which this has for the *coverage*, *timeliness*, *frequency* of updating, *validity*, *reliability* and *consistency* of the derived statistics in relation to the requirements of the descriptive and analytical use to be made of them. The mere fact that the data are recorded as part of an administrative process will have consequences for the quality elements, even if the statistical authority was given the opportunity to influence the content of the administrative reporting system and the rules for its operation. This is a big *if*, and even then the statistical agency will not be able to influence the reporting system's actual operation or the public's reaction to the rules and regulations administered.

2.1. Coverage

Deficiency in coverage may concern the *type of units* actually covered as well as the *degree of coverage*. Hospitals typically register *admissions* rather than *persons*, and cases of illnesses which do not lead to an admission will not be covered; the police registers reported crimes and not all crimes or victims; to take two well-known examples of administrative registrations where the primary unit of observation is an event, but where many users of the statistics would want primarily to focus on, e.g., the person(s) involved. This difference in coverage will in most cases be a consequence of the fact that the primary purpose of any administrative information system is, almost by definition, to reflect the activities of the responsible agency and to serve its management. The nature of the regulations and the attitude of the public to them, as well as the efficiency of the agency and adequacy of its resources, will

also determine the extent to which the *intended* coverage of the registrations corresponds to the *actual* coverage, cf. the discussions of the size and structure of the "grey" economy in many countries.

2.2. Timeliness

The timeliness of data from administrative registrations depends on who is responsible for reporting to the administrative authority and on the incentive for timely reporting. If it is part of the duties of those reporting to observe or participate in the events and activities which are being reported, then the events are more likely to be reported, and to be timely reported, than if they are extraneous to the reporter's normal activities. The "seriousness" or "visibility" of an event, in the eyes of the observer or participant will also influence whether and when it will be reported. Thus the reporting of, e.g., crimes or accidents is normally more timely (and coverage is more complete) the graver the injury or the larger the economic loss. The time needed by the agency to process the report will also significantly influence the timeliness of the statistics which can be derived from them. The production of statistics on household incomes from the information recorded on tax returns may have to wait for the tax assessment process to be completed for all households, even those who choose to dispute the finding of the tax authority.

2.3. Frequency

The frequency whereby statistics can be produced from administrative registrations will depend entirely upon the type of reporting and registration procedures which are being followed by the responsible agency and the way its records are stored. Events are normally reported continuously but they may nevertheless be registered in batches at the end of the week or month. Backlogs in the registration may be carried over to the next period, thus further influencing the timeliness. Once recorded, the frequency whereby statistics can be produced will depend on the cost of processing the records to statistical tables, whether through manual procedures or through the use of computers. (Main-frame processing used to be very costly.)

2.4. Validity, Representation and Coverage of Variables

The administrative agency has to measure the variables which are valid and pertinent for the execution of its duties and which reflect the laws and regulations it implements. These variables and their representation, i.e., the value sets used, may be rather different from those variables and categories which are most valid for statistical description and analysis. One well-known example is that taxable income minus taxes paid may not at all correspond to a concept of disposable income valid for the analysis of income distributions. It is also quite possible that the tax records will not include information about the taxpayer's age, sex, marital status or source of income if such information is irrelevant for the calculation of income taxes.

2.5. Reliability

The reliability of the administrative data depends on the incentives for respondents to give correct (or incorrect) information. Many administrative information systems are linked to schemes where some responses result in a better outcome for the respondent than others. The reliability of the data will then depend on the probability of errors being discovered and on the costs to the respondent – including punishment – if it is discovered that the information given is incorrect. Information on income from tax returns is one obvious example. Another example is information about occupation in schemes for the administration of migrant workers where the rules, specify for example minimum wages by occupation or restrict the recruitment of workers with certain occupations – such as a recent Philippine restriction on the recruitment of “domestic helpers”, which is reputed to have resulted in a sharp increase in the recruitment of “governesses” and female “gardeners” and “drivers”. It should also be noted that even if it is discovered that incorrect information has been recorded it will not necessarily be corrected. If, for example, the placement officers have learnt how to cope with incorrect or imprecise registration of “occupation” on the records of job seekers and vacancies, they may not see the need to correct the records once a satisfactory match has been achieved, thus leaving the record uncorrected for the production of statistics on the occupational distribution of job seekers. These types of errors are likely to be systematic rather than random, with obvious quality consequences for the statistics.

2.6. Consistency

Consistency will have to be evaluated both in space and in time. Many administrative information systems make the registrations in local offices which are under only limited control or instruction from a central office, which therefore may have little possibility of checking the reliability and consistency of the information submitted to it, nor have the authority to order steps necessary to improve data quality if deficiencies are discovered. More serious still is the problem of consistency over time, because so many of the questions asked of official statistics concern what they can tell about changes over time. The most obvious set of problems, and the easiest ones to monitor, if not overcome, relates to changes which are made to formal rules and regulations determining the work of the responsible agency. Such rules may influence *who* should report, *what* should be reported and *how* the reporting should be done, and changes to them have sometimes resulted in the complete disappearance of the basis for important statistical series – cf. the consequences for statistics on imports and exports of the removal of customs declarations for trade between the member countries of the European Union. More insidious are all the small, “informal” modifications in actual reporting and recording procedures which follow from changes in budgets and priorities within the responsible agency. Because servicing the information system is often seen as incidental to the “real” tasks of the organisation – such as helping people in need – servicing the information system may tend to be given lower priority if budgets are reduced or squeezed through uncompensated increases in the workload. This will frequently result in unreported reductions in data quality and may result in systematic changes compared to previously generated

statistics. Unreported improvements in data quality, for example as a consequence of new technology in recording the information, are just as disruptive for the consistency of time-series, at least in principle. In addition to these organisation-related problems, there may be adjustments over time in how the public behaves in relation to the administrative system – adjustments which may reflect an improved understanding of how to use, or misuse, the system rather than reflect a change in the underlying circumstances which the data are intended to reflect. One example is that the number of applicants for certain benefits may increase following news reports or a campaign to increase public awareness of the scheme.

3. How to Use the Administrative Data?

Most official statisticians will agree that in practice it will be impossible to refrain from using administrative data as a basis for official statistics, and that the budgetary and political realities under which national statistical agencies are working and will be working in the future make it likely that an increasing proportion of official statistics will have to be based on administrative registrations. Therefore, the interesting question is what we as statisticians should do to ensure that the official statistics produced on this basis will have the required minimum quality. Fortunately, the answers to this question can be found in the established practices and principles for the production of official statistics:

- Have *detailed knowledge* of the way the observations are being collected;
- *Monitor* the data collection process to ensure consistent data quality;
- Make every effort possible to *improve the data collection* process when it is found to be deficient;
- *Calibrate* the observations generated by the administrative system (through the use of statistical surveys if possible);
- *Estimate* the parameters to be presented in the statistical tables, making use of statistical methods and detailed understanding of the process which have generated the observations on which these estimates are based;
- *Explain* to the user of the statistics their proper use, given the characteristics of the basic data and the methods used to arrive at the statistics.

These points are interrelated and mutually supportive.

3.1. Know the Data Collection Process

The need to know intimately the way the observations have been collected should be obvious, and there is no more reason to expect that actual data collection and processing practice of an administrative agency is more consistent with whatever guidelines exist for its registrations, than there would be with a purely statistical data collection process – which conscientious statisticians will monitor closely, and seek to modify if it does not function properly. The need to know the details should be even more urgent in the case of a process which is under the control of an organisation which, at least partly, has different goals and priorities than those of the statistical agency.

3.2. Monitor the Data Collection over Time

In view of the importance of ensuring consistency in time series, it is of course not enough to establish the particulars of a data collection process at one point in time. The way this process develops over time must be monitored. Such monitoring will have to be partly through information generated internally to the process itself, and partly through the calibration process outlined below. The need to monitor the process may result in a need for information additional to that directly used to produce the statistics – for example: in the case of administrative systems which generate a continuous stream of reports, such as the registration of “unemployed persons” at national employment services, there is a need to know both the date when a person came to the employment service to register and the date this registration was recorded, to be able to monitor how long it takes for all, or a given proportion of, entries/quits occurring in a defined period to be recorded. This information is of vital importance for the decision on when to generate estimates of the number of unemployed persons with reference to a certain date and the number of entrants and quitters during a reference period, and thus for the timeliness of the statistics.

3.3. Help Improve Data Quality

The statistical agency has an obvious obligation to try to help the administrative authorities to improve the quality of the data which they collect, by bringing both individual errors and more general weaknesses to the attention of those responsible for the registrations and by making suggestions for improvements – especially as the experience and competence of the statistical agency with respect to efficient data registration in most cases will be superior to that of the administrative data collectors. However, “data correction procedures across institutional borders are difficult and time consuming” [2], and many agencies tend to deeply resent “outside” interference on their turf, even if the statistical agency has been given an explicit mandate to do so. Even if suggested changes would be seen to result in improved data quality also from the perspective of the data generators, they may not be interested in giving such improvements sufficient priority to actually make the necessary investments and/or changes in rules and/or procedures.

3.4. Calibrate the Results

The natural sciences have a long tradition of calibrating a measurement instrument against other instruments capable of measuring the same phenomenon, to ensure that the performance qualities of the instruments are known and satisfactory. The social sciences, on which official statistics are primarily based, do not have the same tradition. (Crime statistics is a notable exception, cf., e.g., [3].) This is a serious shortcoming which the use of administrative data makes it imperative to rectify. Statistical sampling theory and the accumulated knowledge and experience about statistical data collection instruments can form a basis for the necessary development of methodologies for such calibration on a regular basis. It is likely that the quality of administratively based statistics can only be satisfactorily monitored and calibrated by the use of independent statistical surveys, even when it may be possible to derive closely related statistics from different administrative sources.

3.5. Use the Registrations as Basis for Estimates

When official statistics are based on sample surveys it is elementary that the numbers presented represent estimates of the population parameters. The basis for the estimates is statistical theory, and knowledge about the underlying reality and the way the observations have been generated. The same can be said for the estimated parameters of econometric and similar models. Similarly it is recognised that, e.g., from labour force surveys, the number of unemployed persons should be estimated from responses to questions concerning job search and availability for work. The concepts of employment and unemployment are such that respondents should not be asked directly whether or not they are "unemployed". It should similarly be recognised that the use of administrative registrations to produce official statistics normally will necessitate that they should be used as observations from which one can estimate the parameters which are needed for statistical analysis and description. The estimators to be used need to be based on detailed understanding of the underlying reality which one tries to describe and of the processes which have generated the administrative observations – for example the way and extent to which tax avoidance may bias income estimates – supplemented by statistical methods and insight gained from independent statistical calibration of the administrative data. It is likely that the necessary estimators (and descriptors of their qualities) can only be developed through the increased use of formal structural models in the estimation process – drawing on the methodological experience of econometrics and related fields linked to other social sciences.

3.6. Explain the Results

Official statisticians have an obvious responsibility to explain the proper use which one may make of the statistics which are released, and to warn against unwarranted use of the data. The statisticians should, after all, be better placed than almost all of the users to understand the strength and weaknesses both of the underlying data and of the procedures and methods used to arrive at the released statistics. It is not sufficient to point at weaknesses, such as changes in regulations which cause a break in time-series, and then say that "this should be borne in mind when using the data". This is a cop-out equivalent to saying that the results from a sample survey are subject to sample errors, without saying anything about their size or how they may influence the evaluation of differences between groups or over time.

4. Concluding Remarks

Most official statistics which are not based on household surveys or censuses, do in fact rely on administrative data. The statistical questionnaire is in effect an *indirect use of administrative records* (IDUAR), when asking the person responding on behalf of the enterprise, hospital, school or other organisation to estimate the requested information from the registrations in the administrative records of the responding unit. The more the requested information differs from that available in terms of coverage, definitions, categorisation and/or time reference the less reliable

are the resulting estimates provided by the respondent, the less control does the statistical agency have over how estimates are arrived at and the heavier is the response burden. The increased *direct use of administrative records* (DUAR) as basis for official statistics in effect means that the responsibility for making the required estimates is transferred from the responding to the statistics producing agency. This must have important methodological consequences, even when it is not possible to link records of the same units through the use of *universal (or unique) personal or establishment identification numbers* (UPIN and UEIN respectively) in the way, e.g., done by the Scandinavian statistical offices. In [2] it is said that "In view of the Danish experience it is legitimate to ask whether the problems (in using administrative registration as basis for official statistics) are of such a basic nature that they can only be overcome through systematic development of methodical counter measures." There is no doubt in my mind that the only possible answer to this question is yes.

One of the main challenges now facing not only official statisticians but also the statistical community at large, and the social sciences which support and make use of official statistics, is precisely to find the methodological tools and to develop the models necessary to identify and overcome inherent and procedural flaws of administrative registrations, from the perspective of their use as basis for official statistics. It is likely, however, that the development of such methods and models will demonstrate that one of the main arguments for using administrative data, referred to at the start of this article, namely their relative cost effectiveness, will prove to be wrong for many purposes and in many areas. There will therefore always be the danger that Gresham's "law" will be applicable to official statistics, i.e. that "bad, cheap statistics will tend to drive out adequate, but more expensive statistics." The challenge to official statisticians is to ensure that the applicability of this "law" will be as limited as possible. This challenge is particularly important in countries with weak administrative agencies, where many of the issues referred to in this article will be particularly serious, and where, at the same time, the resources for independent statistics are particularly limited.

References

- [1] Danmarks Statistik, 1994. *Personstatistik i Danmark: Et registerbasert statistiksystem*. Copenhagen, 1994. (An English language edition will be published by Eurostat in 1995.)
- [2] Jensen, P., 1986. Quality Problems Associated with the Statistical Utilization of Administrative Records - Some Danish Experiences. Report submitted by Danmarks Statistik to the UN/ECE/CES Meeting on Statistical Methodology (17-20 February 1986). CES/AC.48/60
- [3] Kester, J.G.C., 1994. Measuring crime: trends and coherence in crime statistics. *Netherlands Official Statistics*, Vol. 9, 1994.

Eivind Hoffmann joined the ILO Bureau of Statistics in 1984 after having worked in Statistics Norway (1968-1981) and the Norwegian Computer Center (1981-1984), in particular with social indicators, labour, regional and environmental statistics and related methodological problems. In the ILO he has been responsible for the revision and follow up of the *International Standard Classification of Occupations* (ISCO-88) and the *International Classification of Status in Employment* (ICSE-93), and he has worked with statistics on employment, unemployment, vacancies and the international migration of workers.