Compilation and presentation of labour statistics based on administrative records

by R. J. Pember[1]

Introduction

Records of labour administration systems hold a wealth of information which are underutilized in many less-developed countries, mainly for lack of statistical expertise. ILO/EASMAT has recently produced a document prepared for use in training labour ministry officials in compiling labour statistics from these records. The document is ILO/EASMAT: *Labour statistics based on administrative records: Guidelines on compilation and presentation*, (Bangkok, 1997).

This publication covering the following topics is summarized in this paper:

- the value of labour statistics derived from administrative records as a component of an overall system of labour statistics;
- the strengths and weaknesses of administrative systems as sources of labour statistics;
- evaluating the statistical value of an administrative system;
- determining which statistical data to extract;
- improving the statistical compilation from a system;
- coding and editing the data in administrative records for tabulation; and
- producing tables and graphs of results.

What are administrative data?

Administrative data are produced as a by-product of the administrative functions of an agency, such as a government department. In undertaking these functions, often under the authority of a set of laws or regulations, an organization will generally register or record a great deal of information which is needed for the administrative process. This information may relate to events or transactions (event-based systems) or to people and organizations, and may be processed and tabulated as statistical outputs.

Statistics compiled in this way may be distinguished from data collected *directly* as part of a statistical census or survey. In direct statistical collections, data are collected for statistical purposes and not in the course of an administrative function. As will be shown later, this difference is important in many ways, particularly in respect of costs and quality:

---

[1] Senior Specialist in Labour Statistics, ILO East Asia Multidisciplinary Advisory Team (EASMAT), Bangkok.

(a) *Costs.* Initial data collection costs for the statistics-producing agency will be less when using administrative records than when collecting similar data through a special survey. Administrative data will be gathered even if they are not used for statistical outputs. Costs other than for data collection may be higher.

(b) *Quality.* Data compiled from administrative records may not conform to normal statistical standards because they are by-products of an administrative process or a legislative requirement which has its own specified standards. An overall evaluation of the relative quality of the resulting statistics will depend on users' priorities.

Administrative data as part of an integrated system of labour statistics

A comprehensive system of labour statistics would include statistics:

(a) from regular and ad hoc household-based surveys and censuses;
(b) from regular and ad hoc establishment-based surveys and censuses, and
(c) as a by-product of administrative systems.

Such a comprehensive system would also require:

(d) legislation and mechanisms for coordinating the statistical programme;
(e) national standards for defining concepts and units;
(f) standards for classifying the data collected, and
(g) the regular maintenance of comprehensive population frames for use in selecting samples for household and establishment surveys. The population frame of establishments may also be used as a basis for labour inspection and other administrative functions (subject to confidentiality constraints).

Statistics based on the administrative records of a ministry of labour (and other agencies) include those on a wide range of subject-matter topics, such as conditions of work, employment of aliens, unemployment, job vacancies, industrial relations and industrial accidents. The statistics are based on a variety of records relating to people, businesses, events, transactions, and so on.

From a statistical point of view, there are major advantages in ensuring that (a) these administrative sources are seen as part of an integrated system of labour statistics, and (b) estimates derived from each administrative system are as compatible as possible with those from household- and establishment-based surveys. Often this can be achieved without deviating from the main priority of the system, that is its administrative purpose. An administrative system is intended to implement the administrative functions of the agency, often in response to legislative requirements or specifications, but this objective can still be achieved while maximizing the use of national standard concepts, definitions and classifications as far as possible to improve comparability and compatibility with statistics from other data sources.

Estimates based on administrative records often differ from those based on data obtained directly from individuals and establishments on the same or related topics. Differences between estimates from

different sources may be of four types - differences in variables, in coverage, in precision, and resulting from measurement errors. These differences can be minimized and results harmonized by appropriate action. (See also Leunis and Altena, 1996.)

Labour statistics based on administrative records may be used in (a) *assessing the productivity and efficiency of an administrative system,* as well as providing a quantitative measure of its coverage and in monitoring the performance of the system, as well as (b) *assessing and monitoring the economic and manpower situation* for use in preparing, evaluating and monitoring action plans, structures and outcomes.

Review of strengths and weaknesses of administrative records as a source of statistics

The compilation of statistics from administrative records has several advantages as follows:

(a) *Low data collection costs* for the statistics-producing agency: Since the data has already been collected as part of an administrative function, there are few costs in accessing the data for statistical compilation.

(b) *No additional response burden* for the respondents: Similarly, the units of enquiry (persons, businesses, other organizations) are spared the inconvenience and cost of a separate statistical enquiry. The data which they have already provided as part of an administrative process (registration, application, inspection, notification) may be used for statistical compilation without them having to be involved in a separate statistical enquiry on the same or closely related topics.

(c) It represents a *full count* of the 'clients' of the administrative system: A complete count is possible since all the records of the system are available for use in the statistical compilation. This means that statistics can be produced for small groups, such as small areas (districts, towns and provinces), without having to be concerned with problems of sampling precision.

However, statistical compilations from administrative systems are not without disadvantages. Possible disadvantages are linked to:

(a) The *types of units* described in the records: The units used in some administrative systems may not be the most appropriate to satisfy statistical user needs (for example, jobs versus persons, establishments versus enterprises) or may not use a definition of the unit which is compatible with other statistical sources. Some systems register persons or organizations, while others register events (which may occur several times for each individual in a given period). Different users may prefer one or the other, but may not be able to extract this information from the system.

(b) The *scope* of the registrations/applications: Some administrative systems may have too narrow a scope in that certain categories of units are excluded by design (legal exemptions) or otherwise (illegal non-registration, avoidance), while others may have too broad a scope in that it may include groups which are not of direct interest to a user. For example, registered job-seekers at

employment exchanges exclude those unemployed who have not registered and may include those employed who are seeking a change of jobs or additional jobs.

(c) The *content of the data* collected: The forms used in an administrative system may not include all the information of interest for statistical users or for statistical processing. The data content may be constrained by legislation, limited resources, or other reasons. Coding of important data may be constrained because the distinctions needed for administrative use are fewer than what users of the statistical descriptions of the units/events will need. Thus the descriptions used as a basis for coding may be incomplete and the system may not use national standard coding classifications. Furthermore, data may be correct at first registration, but not be subsequently updated (see also next point).

(d) The *procedures* for handling data: The procedures used in the administrative systemare designed to serve the administrative objectives, rules and regulations, and not to provide a basis for valid, reliable and timely statistics. Administrative procedures may not require the removal from a data base of expired records, or the updating of job details after a person is first registered. The operators of the system are more likely to edit and correct those data which affect their decision making, administrative action or output, and not to give much attention to other data which do not affect their work but which are important for statistical analysis. Administrative procedures and the flow of forms through a system may also lead to delays in updating a data base.

(e) *High processing costs*: Since administrative procedures are intended to achieve administrative (rather than statistical) output, considerably more attention may need to be given to detecting and correcting errors, and to coding of information which was not needed for the administrative system but is needed for statistical analysis. This processing may require expensive follow up and file amendment.

Evaluating the statistical value of an administrative system

Both the statistician and the statistical user need to be aware of the statistical limitations and constraints of a particular administrative system before using any of its statistical output. The following are some of the issues to be considered when evaluating the statistical value of an administrative system.

(a) *Coverage of units:* What types of units are being recorded (jobs, events, persons, etc)? In theory (according to regulations or legislation), who should be included or excluded from the system? This will be determined by the legislation, operating manuals and procedures of the administrating agency. In practice, what units do the records cover? What incentives are there to encourage registration/reporting to the system? Do many units not register/report? Does the agency have sufficient capacity to record all units, or are some omitted due to lack of resources?

(b) *Range of variables:* What data items are included in the records? Of these, what data items are used by system administrators? What items are coded? With what sort of coding

classification? What definitions and classifications are used? What reference periods do the data relate to?

(c) *Frequency:* How often is it possible to extract statistics from the system? This will depend on the type of reporting:

(1) Continuous reporting of events (e.g. hirings, separations, accidents) can provide statistics with any frequency;

(2) Case-by-case registration (e.g. job seekers, vacancies) can also provide statistics with any frequency;

(3) End-of-period reporting (e.g. of income earned, number of employees) can only provide statistics with the same frequency as the reporting.

Is it possible to extract information at different stages of the administrative process? The system evaluator needs to be aware of the way in which data flows from the time of initial registration to final filing. This includes an analysis of time delays, whether data are added in the process, whether there is any feedback (reverse flow of data), what steps are taken at each stage, how data files are organized and held, and so on.

(d) *Timeliness:* What is the delay between the date of an event, the date of the report, the registration of the report, the processing of the report, the finalization of administrative action on the report, and so on? Statistics compiled from final data sets may be very untimely and it may be useful to consider preliminary statistics based on an earlier stage of the system, if this is possible.

(e) *Geographic specifications:* Does the system cover the whole country, or only urban or only rural areas? Are some provinces excluded? Do the registering offices have well-defined geographic catchment areas? Do the records include a postal address which will give only an approximate location (e.g. town, district) or do they use a precise geographic reference (e.g. a street address or similar)?

(f) *Validity* of variables, definitions and classifications: Are the variables valid and useful for description and analysis beyond the areas of administrative concern? Are the definitions and classifications based on national standards and at least comparable with other data sources? The variables recorded in an administrative system are often for one of three main purposes:

(1) to determine whether the unit is eligible for a service;
(2) to determine the type of services, support or obligation to be provided;
(3) to determine the amount of services, support or obligation to be provided.

These purposes will determine the definitions and classifications used for the variables in the administrative system.

(g) *Reliability* of measurement: The reliability of the data recorded by the system will depend on the following factors:

(1) The incentives for "clients" (those providing data to the system) to give correct or incorrect information;
(2) The cost to "clients" of finding the correct information;
(3) The probability of being found out if incorrect information has been given;
(4) The loss to "clients" if they are found to have given incorrect information;
(5) The cost to the agency of controlling the information received;
(6) The technical means available to ensure the correct recording of information (e.g. for coding occupation);
(7) The gain to the operating agency of correcting wrong information.

The system evaluation should be aware that the quality of data which are collected but not actively used or updated by system administrators are likely to be doubtful.

(h) *Consistency over time:* How stable is the administrative system? Do regular changes occur in the legal exemptions? Are system changes unintentionally introduced as a result of operational changes? For example, do surges in workload lead to unpredictable changes in operational procedures, disruptions in the flow of forms or registrations, or other changes which might affect the comparability of results through time? Changes in the attitude and perceptions of the public will also affect the use and misuse of the system, and hence the quality of the statistics produced by the system.

(i) *Consistency between local agencies:* Are the same procedures, definitions, scope, etc used exactly in all geographic locations? Do different offices apply different procedures depending on their available resources or their use of different operating manuals?

Determining which statistical data to extract

What data items should be used in the statistical compilation? The key issue here is the need to balance costs against benefits. One must achieve a compromise between (a) the minimum data set (and associated resource needs) which will provide sufficient statistics to satisfy basic user needs; and (b) the generally larger data set which users would prefer and which requires additional resources for processing.

Only some of the data items in a particular system will need to be processed and analyzed to produce statistics. The choice of these data items depends on a number of factors, including the following:

(a) What is the minimum data requirement to meet basic user needs? One should distinguish information which is considered "nice to know" from that which is "necessary to know".

(b) What are the costs/resources needed for processing the data items? Data items which are costly to verify or edit or difficult to code are less likely to be always completed and may be less reliable.

(c) What data items do system operators use in decision making? These are likely to be more carefully checked and therefore more reliable than data items not used by administrators/operators of the system.

(d) What data items are not well completed and therefore less reliable? One should examine each item with care and caution. The data may not be as reliable as expected.

The main types[2] of data items which need to be accessed in statistical compilation are:

(a) *identification data.* Each record should have a unique identifier or reference number for ease of filing and retrieving;

(b) *date(s).* The date on which an event occurred, date of registration or reporting of the event, date on which action was taken, date of death (if different from date of accident), and so on, may all be relevant in order to be able to place a record into a particular reference period for tabulation and for calculating durations;

(c) *classificatory data.* Each record should have data to be used in grouping or classifying the record when producing tables or graphs. For example, nationality, occupation and age may be used as classificatory data items;

(d) *quantitative data.* This is the information to be aggregated or averaged. For example, income, age and number of working days lost.

Finally, in selecting data items to be accessed in statistical compilation, one should be aware that the whole process of statistical compilation (including the choice of data items) should be reassessed after an initial period to decide whether costs and quality are appropriate.

Improving the statistical compilation from a system

Statistical compilation from an existing system may be improved as a result of an evaluation of its limitations. Such improvements might include the following:

---

[2] A data item may be both classificatory as well as quantitative. For example, age, income and weeks of employment may be used to classify the results or may be averaged/aggregated.

(a) Existing records may be underutilized and data which were previously unused may be extracted. Data which are missing from the record may be available from other sources and linked to the data record for improved statistical analysis.

(b) The scope and coverage of the system might be improved by amending legislation, introducing (more) incentives and/or disincentives, providing more resources, improving procedures.

(c) The timeliness of output may be improved by extracting data from a different stage of the administrative process.

(d) The speed and reliability of reporting and processing may be improved by redesigning the reporting form. A form which is simple and clear in design and layout may reduce the reporting burden and delays in processing, minimize data errors and omissions, and generally improve the quality of statistical outputs. Key requirements include:

(1) formulating questions in a clear, unambiguous manner;

(2) ensuring that the print is large enough to read;

(3) providing sufficient space for answers to be recorded in detail.

(e) Improved consultation between statisticians, computer system analysts and system administrators. These consultations may improve the overall efficiency of the system as well as the operation and design of any supporting computer system, and thereby lead to improved timeliness, content, reliability, etc. In particular, administrators should be aware of the advantages to their systems of computerization, improved system procedures, improved form design, and standardized classifications.

Before committing oneself to revising or initiating the extraction and compilation of data from an administrative system, it is important to determine the amount and quality of resources available, and what resources are required for the statistical compilation. It may be decided that the cost of these inputs does not justify the statistical outputs achieved. If, however, it is decided to proceed then care must be taken to adequately plan the project implementation, including attention to the timing of inputs, critical paths, approval of budgets, and so on.

Data processing and output

Statistical compilation and presentation will usually require data processing and analysis which:

(a) simplifies details for ease of interpretation (coding);

(b) checks input records for completeness and accuracy (edit/amendment);

(c) summarizes data into tables, graphs and charts (tabulation phase);

(d) interprets and presents the results in a report (reporting).

Coding

Information in administrative records can be used for statistics only if it can be represented in a simplified and formalized manner, i.e. if the various information elements can be coded to a set of defined, relevant categories. Thus the purpose of coding is to classify answers into meaningful groups for subsequent analysis or retrieval. Most administrative records will have codes in respect of some of the information elements on application and registration forms. This coding will have been done to make it easy for clients to furnish the required information and administrative officers to record and interpret it.

Data items may be pre-coded in the form design or coded subsequently. Pre-coding of response alternatives is usually only possible if the number of relevant alternatives is small (say, less than ten) and if it is easy to distinguish between the relevant alternatives. This applies whether the form is to be completed by an interviewer or by the client. Office coding of written responses will be necessary if the number of alternatives is large (say, over ten) and if special training and/or tools are needed to determine the correct group for a response.

Office coding should retain as much information as possible from the response. Consequently, it is recommended that the responses should be coded to the most detailed level of the classification supported by the information given and the specified coding rules. Vague responses should be coded to the appropriate broad categories of the classification. Office coders should have the required tools for this work, including:

(a) a coding index which is organized to support coding rules on matching a response to the correct index entry;

(b) a manual providing the coding rules and clear procedures for resolving queries for responses not reflected in the coding index. The coding index should also be updated regularly using the results of these queries;

(c) adequate training of the coders.

The coding classifications used in administrative systems have to support the implementation of relevant laws and regulations. If possible, they should also be compatible with national standard classifications, because:

(a) comparison with results from other sources may be easier;

(b) supporting materials (alphabetical and numerical indexes and coding instructions or manuals) may already be available;

(c) coding staff can more easily be transferred from other sections or agencies without having to retrain them.

If there is no national standard classification for a particular topic, it may be useful to ensure that the classifications used are compatible with the relevant international standard classification.

Data editing and correction

The data in the administrative records (or coded from these records) should be checked for completeness and accuracy because errors can be introduced into the records in many ways:

(a) during reporting (either deliberately by the "client" or the officer, or through misunderstanding);

(b) when registering the reported information. The person recording the response may tick the wrong box by mistake, or may rephrase the reported answer in such a way that the recorded information differs from the reported information;

(c) when coding textual information. Coders may misclassify a recorded answer because of tiredness, laziness or misunderstanding;

(d) when transcribing recorded information into computer-readable form. Errors may occur in data entry because of misunderstood handwriting or tiredness.

The risk of errors can be minimized by regular spot checks to verify reporting, careful supervision, suitable staff training and motivation (including the supply of good working conditions, reasonable deadlines, appropriate work instruments and a motivating work environment), careful attention to form design and operation manuals, and so on.

Computerized editing will detect many (but not all) errors. It will find values which are outside the legal range (including those which are not possible, not reasonable or missing) as well as those values which are inconsistent with other recorded data (for example, a male receiving a maternity benefit). However, data which are reasonable and/or consistent (but wrong) will not be readily detected.

The tools of error detection should also include check coding, check editing and the use of check digits in unit records, and simple output editing of aggregated results. Check coding is an edit process in which records are recoded and new codes compared with previously determined ones. Check entry is a similar process in which the data capture process is repeated for some or all records. Check digits may be used in a system to minimize errors, especially with key data such as identification numbers where an error would prevent or complicate data amendment. Output editing involves assessing the outputs to ensure that they appear reasonable and suitable for publication.

Archives

Finally, in respect of data processing, one should not overlook the importance of proper storage of data records. This includes sorting paper records by reference dates and identification number and packing them in small, clearly marked piles to ensure easy retrieval. Paper records and computer files should be stored in locations which are secure against theft, physical damage (fire, flooding, humidity)

and natural catastrophes (earthquakes and cyclones). Computer files should also be backed-up regularly on a separate disk and transferred to new media when equipment or systems are changed.

Statistical outputs

The output phase of a statistical system converts data into tables, graphs and other pictorial forms, interprets the results and presents them in a report. Statistics presented in the tables or graphs may be absolute numbers indicating a level or a change in levels, or averages, ratios, rates of change, and so on. The statistics may relate to:

(a) *events during a period*, such as a month or year (for example, the number of work permit cards issued during 1996); or

(b) the *stock of units at a particular point in time*, such as the end of a month or year (for example, the number of current work permits at 31 December 1996).

Both measures (events and stocks) are important for analysis.

The design and preparation of statistical tables require experience and practice. Some of the features of good table design and preparation are mentioned below:

(a) The table should be based on a *complete set of records* for the topic being tabulated. To achieve this some imputation of missing data may be required based on previous reports for the same units or on information from other sources. If details of the missing data are expected in due course, the table should indicate that the results are provisional or preliminary and subject to change when complete results are available.

(b) The table should relate to a particular *time reference period* or a set of such periods.

(c) The table should have a *title* stating what is being tabulated (that is, the units), what classifications are involved and how they are used, and the reference period. The title should be brief but clear and self-explanatory.

(d) The table should include *footnotes* which clarify the concepts used, highlight special important features affecting the statistics and provide detailed definitions.

(e) It should show the *unit of measurement* being used, unless this is obvious from the title. The unit of measurement may be given in brackets between the title and the table itself (if the same unit applies throughout the table) or in the column or row headings (if different units are used in different columns/rows).

(f) It should indicate *totals* for the components of the classification(s) used. Some collapsing of minor classification groups may be necessary for ease of presentation and interpretation.

(g) The table should *fit widthwise onto a page*, even if landscape presentation or collapsing the widthwise classification is required. Its length may extend to more than one page, but the title and

column headings should be repeated on subsequent pages. A table which does not fit widthwise onto a page is difficult to read and may confuse users, particularly since the total of columns will apparently not be the sum of column components.

(h) The *table classification groupings* should be as compatible as possible with those used in other sources (for example, for age groups "under 15 years", "15 to under 25 years", and so on) to facilitate comparison with other sources.

(i) Table specifications should, as far as possible, *anticipate user demands* which have not yet been articulated. Alternatively, the compilation system should be flexible enough to permit new tables to be produced later without major reprocessing.

Tables may be produced monthly, quarterly, annually or at ad hoc periods according to the frequency of data availability, user needs, and costs and available resources.

The timing of producing tables and publications is obviously related. For example, quarterly publications would not be feasible if tables can only be produced annually. The publications programme should take this into consideration.

Graphs, diagrams and other pictorials are essential tools for the presentation of results and the interpretation of statistical data. Graphs may be linear, bar charts, pie charts, and so on. Pictorial displays have a strong visual impact but are not always accurate measures of the quantities involved. This imprecision can be misleading. Care should be taken to select the most appropriate type of presentation.

As with table design, a graph should be self-explanatory. The titles of the graph and axes should be clear and unambiguous. The graph title should state the reference period(s), item(s) being graphed, classificatory variables used and geographical region. The axes: titles should state the item being graphed and unit of measurement. Sources of data might be given in a footnote. Graphs and diagrams should be supported by textual comment to interpret and explain them to users.

Report preparation should be planned well in advance. Table and graph outlines, broad textual comments and annexes should be prepared in advance of table production so that statistics can be inserted and the publication prepared for release without delay. Computer software for desktop publishing, combined with spreadsheet and database analysis packages, facilitate the preparation of high-quality outputs which may be photocopied or sent to printers in camera-ready form, depending on the time and resources available.

Conclusion

There is growing interest world-wide in improving the use of administrative data as a source of labour and other statistics. Administrative records provide one of the three major sources of statistical information (along with direct statistical collections from households and businesses) and are likely to be cost-effective sources despite the limitations mentioned above.

This paper summarizes a document of over 100 pages and clearly omits a great deal of detail and qualifying remarks. Extra information is available in ILO/EASMAT (1997) available from ILO=s East Asia Multidisciplinary Advisory Team, PO Box 2-349, Rajdamnern Bangkok 10200 Thailand.

ILO/EASMAT (1997) was used as the main training reference in the *ILO/Japan Regional Training Workshop on improving labour statistics derived from administrative records* held in Pattaya, Thailand, 17-21 February 1997 with funding by the ILO/Japan project *Assistance to ministries of labour on improving labour statistics derived from administrative records*. It has subsequently been translated into Mongolian, Thai and
Vietnamese for use in national training seminars and for wider distribution.

References:

Brackstone, G.J.: AIssues in the use of administrative records for statistical purposes@, in *Survey Methodology* (Statistics Canada) June 1987, Volume 13, Number 1.

Harala, R. and Reinikainen, A-L.: "Confidentiality in the use of administrative data sources", in *Statistical Journal of the United Nations ECE 13* (1996).

Hoffmann, E.: AWe must use administrative data for official statistics - but how should we use them?@, in *Statistical Journal of the United Nations ECE 12* (1995).

Hoffmann, E. and Lawrence, S.: *Statistics on international labour migration: A review of sources and methodological issues*, Interdepartmental Project on Migrant Workers 1994-95 (Geneva, ILO, 1996).

Leunis, W.P. and Altena, J.W.: ALabour accounts in the Netherlands 1987-1993", in *International Statistical Review*, Vol. 64, No. 1, April 1996.

ILO/EASMAT: *Labour statistics based on administrative records: Guidelines on compilation and presentation* (Bangkok, 1997).