Journal of Official Statistics
Vol. 3, No. 1, 1987, pp. 83–91

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Statistics Canada's Policy on Informing Users of Data Quality and Methodology

Statistics Canada[1]

Statistical agencies often stress the need for users who analyze or interpret a data set to take account of both the quality of the data and of the methodology used to produce the data. Information on the underlying concepts, definitions, data quality, and methodology is essential for determining the relevance of a particular data set to a particular purpose. Information on these points is essential to distinguish real phenomena from random variations or inherent biases.

It follows that the statistical agency is obliged to make this information freely available to users of its data. Given the inherent difficulty for external clients to assess *directly* the quality of statistical information (which, in most instances, verges on the impossible), it is often the statistical agency alone that can provide indications of data quality. Furthermore, it is the producer of the data, the statistical agency, who is in the best position to describe the methodology used to produce the data and any limitations in that methodology.

At Statistics Canada, the traditional recognition of this obligation was first expressed in a 1978 policy statement. This policy statement required the agency's statistical products to be accompanied by a description of the methodology and indications of data quality. At that time some surveys (e.g., the Census of Population and the Labour Force Survey) already followed this practice, but the majority of surveys did not. This policy caused many programs to examine their statistical outputs and comply with the explicit policy. The Publications Board (an internal body charged with central approval of publications) was made responsible for ensuring conformity to the policy. When issued in 1978, this policy had the desired effect of drawing managers' attention to the need to produce and disseminate quality information. The initial effect has, however, deteriorated over time. In 1983, a general review of Statistics Canada's policies

[1] Mr. Gordon J. Brackstone, Assistant Chief Statistician, Informatics and Methodology, Statistics Canada, Ottawa K1A 0T6, Canada, can be contacted for further information on this topic.

and procedures showed that the implementation of this policy required management attention. Indeed, the monitoring mechanism, the Publications Board, had been abolished in the interim.

One of Statistics Canada's internal management committees was charged with reviewing the 1978 policy and recommending appropriate changes. A task force that included survey managers, data analysts, and survey methodologists was created to conduct the detailed review. In particular, they looked at the feasibility of these policy goals, since an important reason for the previous lack of adherence was the high cost of or real impossibility of conforming to these guidelines. They also considered the introduction of an effective method for ensuring adherence. Another objective was to extend the policy's applicability to all statistical outputs, as opposed to the narrower focus on survey-based data that was inherent in the 1978 policy. It was recognized that the policy's extension to data derived from administrative records and data derived from analytic manipulation of data from different sources would present new challenges in developing data quality indicators.

The result of the task force's work was carefully reviewed by Statistics Canada's senior management, and by its external Advisory Committee on Statistical Methods. The main issue faced in revising the policy was how to frame the policy statement so that it would reflect the underlying intent while still taking account of the wide range of the agency's statistical products. Any policy adopted must also consider the reality of limited budgets, and most important of all, the intrinsic difficulty brought about by the virtual absence of *comprehensive* theoretical guidance regarding non-sampling errors and the paucity of rigorous empirical work. One had to acknowledge that the amount of effort and resources that should be put into data quality measurement would depend on the importance and uses of the statistical output and on the program's budget. Furthermore a survey manager would often be faced with a trade-off between adding to the content, or putting more resources into measuring data quality.

Not surprisingly, we could not formulate general solutions to these trade-offs. After long debate, the resolution was to formulate the policy as a goal towards which programs should aspire, taking into account the use of their output, their budgetary restraints, the intrinsic difficulties involved in developing indicators of quality, and their other priorities. Program managers' proposals for bringing their outputs into conformity with the policy must compete with proposals for other types of activities. This process leaves the program manager with the responsibility of recommending priorities for expenditures on his/her program, while allowing the agency's corporate management to determine the relative importance to be attached to this policy goal at the time of resource allocation. An important aspect of the new policy is the requirement for a periodic report on the current compliance with the policy. This report allows an assessment of the speed with which the agency is moving towards its goal.

*

Statistics Canada's Policy on Informing Users of Data Quality and Methodology[2]

Introduction

The measurement of data quality is a complex undertaking. There are several dimensions to the concept of quality, many potential sources of error and typically no comprehensive measure(s) of data quality. A rigid requirement for comprehensive data quality measurement for all bureau products is not achievable given the present state of knowledge. Nevertheless the bureau has an obligation to make its users aware of at least the major describable or quantifiable elements of quality, and of the methods that underly the data being published. Therefore this policy represents a goal toward which all programs should head, while recognizing that budgetary and other constraints may prevent the full attainment of this goal.

Policy

1. Statistics Canada will make available to users indicators of the quality of data it disseminates and descriptions of the underlying concepts, definitions and methods.
2. Statistical products will be accompanied by or make reference to documentation on quality and methodology.
3. Documentation on quality and methodology will conform to such guidelines as shall from time-to-time be issued.
4. Exemption from the requirements of this policy may be sought in special circumstances using procedures described under Responsibilities below.

[2] This policy statement and its accompanying guidelines are unabridged reprints from a Statistics Canada memorandum issued in April 1986.

Scope

1. This policy applies to all data disseminated by Statistics Canada whether collected, or merely assembled, by Statistics Canada. In the latter case, documentation should also describe the particular role of Statistics Canada in the production of the data.
2. This policy applies to all data disseminated outside Statistics Canada through any medium (CANSIM, print, computer tape, micro-film, diskette, etc.) and to any class of user (federal departments, provinces, general public, etc.). The method of making available to users information about the quality of data and about the existence of documentation on methodology may, of course, vary depending on the nature of the dissemination medium.
3. This policy applies to all data disseminated by Statistics Canada however funded. Sponsors of surveys reviewed under the Federal Government Information Collection Policy will be encouraged to conform to this policy.
4. This policy applies to data produced or disseminated in the course of analysis.

Definitions

Indicators of Data Quality

Indicators of data quality are measures or descriptions which summarize the likely magnitude and important sources of differences between the published data and the quantities that the statistical activity was designed to estimate.

Description of the conceptual framework, definitions, methods, external influences on and other features of the data, may also be relevant to the user's assessment of the suitability of the data for particular purposes.

Documentation on methodology

Documentation on methodology is the description of the underlying concepts, definitions and methods used in the production of the data.

Responsibilities

Program Areas will be responsible for:
- the dissemination of existing measures or descriptions of data quality and documentation on methodology;
- the addition of procedures to generate information on data quality, if not already part of the program;
- the inclusion of requirements to satisfy this policy in the design, schedule and budget of new or re-designed surveys or data integration activities; and
- the submission to the Methods and Standards Committee of applications for exemption from the requirements of this policy.

The Methods and Standards Committee will be responsible for:
- the production of periodic reports on the state of compliance with this policy;
- the initiation of periodic evaluations of the application of this policy within particular program areas and ensuring that such evaluations are co-ordinated with quality assurance and program evaluation exercises;
- the provision of guidelines on the application of the policy to program areas;
- the initiation of a review of the policy and accompanying guidelines when deemed necessary; and
- the review and approval of applications for exemption from the policy requirements.

Inquiries

Inquiries relating to the interpretation of this policy should be addressed to the chairperson(s) of the Methods and Standards Committee.

Guidelines on the Documentation of Data Quality and Methodology

Introduction

1. Statistics Canada, as a professional agency in charge of producing official statistics, has the responsibility to inform users of the concepts and methodology used in collecting and processing its data, the quality of the data it produces, and other features of the data which may affect their use or interpretation.
2. Data users have to be able to verify that the concepts they have in mind are the same as, or sufficiently close to, those employed in collecting the data. To do this, a knowledge of the underlying conceptual framework and definitions used in the data collection is required.
3. Users generally recognize that data are subject to error and therefore need to know whether the data are sufficiently accurate to be useful to them. To make this assessment, they need to be informed of the likely principal sources of error and, where possible, the size of the error. They also need to know of unusual circumstances which might influence the data.
4. Given that indicators of data quality cannot be expected to be comprehensive, data users also require a knowledge of the data collection and processing methodology in order to verify whether the data adequately approximate what they wish to measure and, whether the estimates they wish to use were produced with tolerances acceptable for their intended purpose.

The Guidelines

5. These guidelines are primarily intended for internal use at Statistics Canada when the documentation and dissemination related to a statistical programme are being planned or reviewed.
6. The level of detail to be provided in documentation on data quality or methodology will depend to a considerable extent on the type of data collection, the medium of dissemination, the range and impact of uses of the data, and the total budget of the collection/production process. Managerial discretion is required in determining the level of detail appropriate for a given data set.

Guidelines on the description of data quality

7. Data quality is generally described in terms of sampling and non-sampling errors.
8. Unexpected events which influence the data should be flagged for users to help them in interpretation of the data.
9. It is not generally possible to provide comprehensive measures of data quality. Rather one should aim to identify what are thought to be the most important sources of error and provide quantitative measures where possible or qualitative descriptions otherwise. The result should be a balanced discussion which addresses itself to specific sources of error or bias and is therefore informative to users.
10. For censuses, surveys or administrative data surveys, the description should cover as many as possible of the elements described in paragraph 14 below.
11. Index numbers of product prices or quantities can be treated similarly, but their conceptual basis presents an additional dimension necessary to describe data quality. Particular attention might be given to any substitutions made in developing the estimates with special reference to product changes and changes in product quality. (For the corresponding methodology description, see paragraph 20 below.)
12. In the case of national accounts and data resulting from analytic activities, both the impact of quality problems in the source data, and the impact of the methods of analysis, integration, benchmarking and adjustments used, have to be taken into account. (For the corresponding methodology description, see paragraph 21 below.)
13. Statistics derived from administrative data or data not collected by Statistics Canada can also be dealt with under the guidelines of paragraph 14, but it is likely that the information available will be less detailed. Nevertheless, important issues such as coverage, response errors and comparability over time should be discussed.

Data quality descriptions

14. The following aspects of data quality are regarded as basic and, subject to constraints of cost and feasibility, some indication of their level should be provided or made available, where applicable, for every statistical product:
    a) *Coverage* – the quality of the survey frame or list (for surveys or censuses) or source files (for administrative data) as a proxy for the desired universe should be addressed (including gaps, duplications and definitional problems).
    b) *Sampling error* – if the survey is based on a random sample then estimates of the standard error of tabulated data based on the sample should be provided, together with an explanation of how these standard error figures should be used to interpret the data. The

method of presentation may vary from explicit estimates of standard errors to use of generalized tables, graphs or other indicators. If the survey is based on a non-random sample, the implications of this on inferences that might be made from the survey should be stated.

c) *Response rates* – the percentage of the target sample or population from which responses or usable data were obtained (on a question by question basis if appropriate) should be provided. Any known differences in the characteristics of respondents and non-respondents should also be described as well as a brief indication of the method of imputation or estimation used to compensate for non-response.

d) *Comparability over time* – it may be appropriate to discuss comparability with the results of the same activity for a previous reference period, especially if there has been a change in methodology, concepts, or definitions. If such a change would affect comparability from one time period to another, a quantitative estimate of this effect should be made whenever possible.

e) *Benchmarking and revisions* – the effects of benchmarking or revisions on comparability over time should be described. Guidance on the possible impact of future benchmarking should be given based on past experience.

f) *Comparability with other data sources* – if similar data from other sources exist, they should be identified. Where appropriate, a reconciliation should be attempted describing how the data sets differ and the reasons for these differences. Comments on quality of the other data should be provided if an evaluation is available.

g) *Other important aspects* – there may be other aspects of data quality that are of prime importance given the objectives of a specific activity. These should be included with the basic indicators of data quality. Examples are: unusual collection problems, misunderstandings of the intended concepts by respondents, major strikes, changes in classification or in its application, response based on financial years that do not correspond to a fixed reference period.

In larger repeated surveys or activities, the most recent available information on more detailed aspects of data quality may also be provided in separate reports on data reliability. Such quality measures will usually be derived as the results of special evaluations. In different surveys and at different levels of aggregation, different sources of error may predominate. Subject to cost limitations, the most important sources of error should be evaluated periodically, and the results made available to users in the most convenient form.

h) *Total variance (or total standard error) and/or its components by source* – the overall variability of the statistics, including the effect of sampling error, response error, and processing error, should be provided if an appropriate model to aggregate these sources of error can be constructed at a cost which is reasonable relative to total program budget.

i) *Non-response bias* – an assessment of the effect of non-response on the statistics should be provided if possible.

j) *Response bias* – evidence of response bias problems stemming from respondent misunderstanding, questionnaire problems, or other sources, should be provided if available.

k) *Edit and imputation effect* – the effect of editing and imputation on the quality of data should be assessed.

l) *Seasonal adjustment* – measures of the impact and significance of the adjustment should be provided together with an explanation of how these measures should be interpreted. Examples of such measures are the mean absolute percent change of the last year's revisions of the seasonal factor, or the MCD (months for cyclical dominance) statistic.

m) *Any other error sources* – if there are particular sources of error or unforeseen events which are relevant to the series or occasion, these should be described.

Guidelines on the description of methodology

15. While all users should be provided with some appreciation of the methodology, some will require greater detail. Therefore, two levels of documentation should ideally be available:
    a) general user reports that are prepared for a wide audience in order to assist them in interpreting the data and in deciding on their appropriateness for a particular purpose;
    b) technical reports that are definitive and exhaustive and give full and detailed information on methods underlying the data.
16. The amount of detail covered in methodology documentation will vary with the type of data collection (census, sample survey, administrative data survey, index, national accounts), the medium of dissemination, the range and impact of uses of the data, and the total budget of the program. A reference to available documentation may be sufficient, especially when the dissemination medium is a short response to a special request, or a summary report.
17. For data resulting from Statistics Canada surveys or censuses, the methodology reports should provide at least an outline of the main steps in conducting the survey and should provide more detailed information on those aspects of survey methodology which have a direct impact on the quality and applicability of the data produced from the survey. The following topics should be covered where applicable.
    a) objectives of the survey;
    b) the target universe and any differences between this and the survey frame actually used;
    c) the questionnaire(s) used and all important concepts and definitions (the discussion of concepts and definitions may well be very lengthy and require a separate document, or they may be included with the survey results; in the former case a reference to the separate document should be made);
    d) the sample design and estimation procedures;
    e) the method used for collecting the data (e.g., interview, telephone, mail, etc.) and details of any follow-up procedures for non-respondents;
    f) any manual processing (e.g., coding) that takes place prior to data capture;
    g) the method of data capture;
    h) quality control procedures used in connection with operations (e)–(g) above;
    i) procedures for editing the data and for handling non-response and invalid data;
    j) benchmarking and revision procedures used;
    k) seasonal adjustment methodology used;
    l) the form in which the final data are stored and the tabulation or retrieval system, including confidentiality protection procedures;

m) a brief summary of the results of any evaluation programs; and
n) any other special procedures or steps that might be relevant to users of the survey data.

18. For statistical data that are collected (often cooperatively) by agencies other than Statistics Canada, but which are assembled and published as a Statistics Canada product, the methodology report should cover the same points as in paragraph 17 to the extent possible, making a careful distinction between those activities for which the collecting agency is responsible and those for which Statistics Canada is responsible.

19. For administrative data, the original purpose of collection is not generally the same as that for which they are used by Statistics Canada. Although the information described in paragraphs 17 and 18 is desirable, the following topics should at least be covered:
    a) the data sources;
    b) the purposes for which the data were originally collected;
    c) the merits and shortcomings of the data for the statistical purpose for which they are being used (e.g., in terms of conceptual and coverage biases);
    d) how the data are processed after being received and what, if anything, is done to correct problems in the original data set; and
    e) the reliability of the estimates, including caveats where necessary.

    It will be noted that items (c) and (e) should be covered in a discussion of quality.

20. For index numbers which are based on data collected through specific surveys or derived from administrative or other sources, the corresponding guidelines in paragraphs 17–19 apply. This applies to data on the main variable (i.e., prices in the case of a price index and quantities in the case of a volume index) and on the index weights.

    In addition, particular attention should be paid to specific conceptual and methodological aspects of index making. Their proper description, in many cases, may be more important for users than a strict assessment of the quality of input data. The following elements should be developed:
    a) *Precise definition of the underlying economic concepts that the index numbers are intended to measure.* Reference should be made to any application or class of applications (e.g., deflation of macro-economic aggregates) for which the index numbers are not suitable.
    b) *The methodology adopted.* This should cover topics such as the index formula, weighting system, computation of the index at various aggregation levels, basing, re-basing, linking of indices, treatment of changes in the varieties or qualities of goods available on the market. The adopted methodology should be compared with the underlying index concepts and possible distortions discussed.

21. Methodology reports for data resulting from analytic activities (including the System of National Accounts) should cover the following topics:
    a) the conceptual framework for the analysis (e.g., the system of national accounts);
    b) the major definitions and concepts used and how they are defined operationally;
    c) the data sources used, and the extent to which they measure the target concepts, as well as gaps and deficiencies in these data sources. Non-compara-

bility of data elements available from different sources should be noted. For analytical activities, reference should be made to the quality of the primary data underlying the analysis;

d) the methods used in integrating and analysing the data from feeder sources including, where relevant, the adjustments made to data from different sources, the methods used for price deflation, the methods used for seasonal adjustment, and/or benchmarking, and a description of the revision process; and

e) any discrepancy arising in the integration or analysis of data from different sources, and the procedures by which these discrepancies were handled (e.g., the statistical discrepancy arising in the estimation of income and expenditure accounts).

March 12, 1986

Received November 1986
Revised February 1987