

# DATA RELEASE CRITERIA IN OFFICIAL STATISTICS

Gordon Brackstone (*bracgor@statcan.ca*) and Richard Burgess (*burgric@statcan.ca*)

Statistics Canada

120 Parkdale Ave.

Ottawa, Ontario, Canada K1A 0T6

## 1. Introduction

Legislation governing national statistical offices (NSOs) usually specifies confidentiality requirements constraining the dissemination of statistical information. These requirements typically lead to the suppression of some figures because of their potential to disclose information about an identifiable individual or unit. No equivalent legislative provisions govern the suppression of information because of its inaccuracy. Statistical agencies have to decide whether and when they will suppress or withhold inaccurate statistical estimates. This paper examines principles that might govern such decisions in a NSO.

To illustrate the dilemma that NSOs face, consider two alternative views, neither completely without merit. The first view asserts that NSOs have no business deciding which data are good enough for users and which are not. They should publish (or make available) all the data they can (subject to confidentiality restrictions), together with descriptions of its accuracy and other relevant characteristics, so that users can make their own decisions about what is good enough for their particular purposes. At the other end of the spectrum is the view that NSOs have an obligation to protect users, the public, and the NSO itself, from the consequences of the use of misleading statistics. By releasing data of questionable accuracy the NSO risks undermining its reputation, and thereby damaging future confidence in its outputs. Where on this spectrum should an NSO aim to be?

## 2. Data accuracy and program design

Data accuracy<sup>1</sup> considerations cannot be divorced from the purposes for which the data are to be used. Data may be quite good enough for one purpose, while being totally inadequate to meet another purpose. Furthermore, a NSO cannot pretend to predict all of the uses that might be made of its data, either when it is designing its programs or when it is deciding which aggregates to publish. No absolute threshold of acceptable data accuracy can or should be defined across all programs. Even within programs, defining what constitutes an acceptable level of accuracy is not a simple task. Nevertheless, each program should be able to define accuracy levels, or *design standards*, that it is attempting to achieve for certain key variables. These design standards may evolve during program design as cost, accuracy and other considerations are balanced. Realistically, design standards can only be defined in terms of the impact of a few recognized and measurable sources of error on a few principal statistics. The levels of accuracy actually achieved will vary because (i) not all statistics are subject to design standards, (ii) for those that are, other sources of error may have arisen as the data were produced, or (iii) accuracy levels achieved vary from those expected. A range of data accuracy levels will always exist among the statistics produced from a single program.

## 3. Principles

Within each program we can think of data in three broad categories in terms of how fit for use we feel they are. This categorization could be based on a combination of explicit measures of accuracy,

---

<sup>1</sup> In Statistics Canada's terminology, *data quality* is defined as fitness for use; accuracy is one of its six dimensions. The others are: relevance, timeliness, accessibility; interpretability, and coherence.

indirect indicators, and subjective views. First, there are data that we gladly stand behind as fit for most uses. These we readily release accompanied by indicators of accuracy. Secondly, there are data that we think can be useful for some purposes if used with care. To these we would like to attach warnings as well as accuracy measures. Finally, there are data that we would rather not release because we feel they could be misleading or widely misused.

By reference to these three categories, six principles are proposed as a basis for managing dissemination decisions in the face of accuracy considerations:

- X each statistical program should establish broad criteria for distinguishing between data in the three categories;
- X pre-planned products should be designed to contain primarily data in the first category;
- X whenever data in the second category are included in pre-planned products or otherwise released they should be flagged;
- X data in the third category should not be included in statistical products and should only be released if a user insists after being fully informed and warned;
- X only data in the first category should be actively marketed;
- X data accuracy and methodology information should be provided with all products.

The above points encapsulate broad intent but leave considerable judgement to programs. Pre-planned products are distinguished from customized requests because, in the latter case, there is the chance to interact with the user when the data are requested. Special rules are required for products intended as sources of aggregation (e.g. public microdata files), for analytic products, and for customized requests.

#### **4. Application**

The application of these principles depends crucially on programs' ability to differentiate the three categories of data from their programs. *De facto*, programs are making dissemination decisions. The challenge is to identify the criteria being used, and then to see that they do not vary unjustifiably between programs. The approach has been tested on a few programs at Statistics Canada. Examples of criteria include the following:

- X a business survey uses a two-way table of sampling error (CV) against imputation rate; each cell is identified with one of the three categories;
- X a household survey has designed its sample and standard publications so that national and provincial data are only suppressed for confidentiality reasons; data with higher CVs fall in category 2; for estimates for smaller areas (e.g. cities), a CV cutoff defines category 3;
- X in all cases, programs leave open the possibility that some estimates may be put in the second or third category if unforeseen response, nonresponse or processing problems arise in the survey.

#### **5. Conclusion**

A common management approach to release decisions is required within a NSO, but not necessarily common accuracy criteria. Programs are so diverse, and user requirements so wide-ranging, that a common accuracy standard should not be imposed. However, neither should release criteria vary capriciously from program to program. In defining criteria, we suggest an approach that uses category 3 primarily for data where serious bias is suspected, and allows high variance estimates to remain in category 2.

#### **Résumé**

Ce document décrit les stratégies de diffusion ou de non-diffusion des données selon leur exactitude. En classant les données de chaque programme statistique selon leur applicabilité, le document propose des principes qui devraient régir la conception des produits et la diffusion de données par les organismes

statistiques gouvernementaux.