Statistics Division
United Nations

Department of Statistics
Malaysia

Seminar on
Data Dissemination: Emerging Trends and Issues
1- 3 August 2007
Kuala Lumpur, Malaysia

Providing Access to Microdata:
A Perspective from Statistics South Africa

Elizabeth Gavin, Nireen Naidoo and Edroy Christians

Statistics South Africa

August 2007

Abstract

The users of statistical information may be segmented into different groups with different requirements. While some users may rely on a single index, others require the ability to explore the statistical data in more depth. Subject specialists require specialized tabulations for research and analysis, which may not be covered in statistical releases or standard data products. The latter group are the primary users of micro-data, and at times their engagement with the statistics agency becomes fraught, as the agency works to protect confidentiality of respondents, whilst at the same time being responsive to users' needs and specialized requirements. Increasingly users are requesting from Stats SA the ability to track respondents over time, which, apart from technical constraints, raises further confidentiality constraints. As the statistical agency within a relatively young democracy, Stats SA has to take very seriously both the need to educate and reassure respondents that information they provide will be kept absolutely confidential, as well as satisfy users that they are being provided with the most accurate and complete information possible, given that the historical context facilitated suspicion about the use of information by government.

Stats SA has been making available confidentialised unit records to specialist users. Increasingly, statistical data are also packaged, so that users can create cross-tabulations of their choice, without directly accessing the microdata. Stats SA has also in special cases permitted use by researchers of microdata, under strict usage conditions. As requests for and the distribution of statistical releases and information products are tracked over time, this provides insight into improvements that can be made to statistical products provided, in order to meet user needs while ensuring respondent confidentiality, and in a way that limits the number of specialized requests, which are time-intensive to fulfill. Stats SA is also focusing on the recording and provision of improved metadata on its data collections and products, which will add value to the information provided, through aiding its interpretation and the assessment of comparability with other information sources.

Introduction: meeting a variety of statistical data needs

1. Statistics South Africa (Stats SA) provides a wide range of statistical products and services, as it recognizes that different users have different requirements in terms of format and the level of detail of statistical information which will best suit their needs.

2. A scrolling banner on the web-site home page[1] – and on Stats SA's head office building – provide key indicators such as the Consumer Price Index (CPI), Gross Domestic Product (GDP), unemployment rated and the national population figure for those who simply need these at a glance. These indicators, together with additional indicators, are also available in printable format from a web-page, through a single click on the home page. The home page also leads - through both the publication schedule and the publications search facility - to all Stats SA's published statistical releases and reports in PDF, which encompass key findings for that particular publication, narrative and graphs and tables as appropriate. Users may also access time series datasets from the website and have access to tools to perform cross-tabulations, without directly accessing the underlying microdata. Microdata is also provided to users; while some is available for downloading through the website, in general the size of the datasets and the limitations on internet bandwidth within South Africa mean that the preferred medium for distribution is via CD.

3. Key to deciding on the statistical products and services needed, including the provision of access to microdata, is the consultation of users themselves. The following section describes interactions with users – and respondents - which guide Stats SA in its development of and improvement upon statistical products and services. The next section provides some examples of the microdata provided by Stats SA. Comments on future trends conclude the paper.

Opinions and perceptions of Stats SA's users and respondents

[1] See http://www.statssa.gov.za

4. Stats SA monitors user opinion on its services and products both directly, through surveys and consultation, and indirectly, through monitoring the access of statistical releases and data through the website and user requests.

5. Stats SA segments users into various groupings, as different users have different needs with respect to content, as well as level of detail. For instance, the information needs of government officials at national, provincial and local level differ. Likewise, private sector users of statistical data are likely to have different needs (e.g. shortest possible lag times between reference period of statistics and publication dates) to academics and researchers (e.g. times series comparable over long periods). These different needs become apparent through direct interactions with users.

6. In a series of nine workshops, one held in each of South Africa's provinces during October/November 2006, two statistical releases were discussed with workshop participants, with a view to obtaining inputs from users as to how these releases might be improved upon to communicate better with users. Participants in these stakeholder workshops spanned all spheres of government, the private sector, universities, non-governmental organizations, international organizations and the media. One of the publications discussed at these workshops was the monthly release on the CPI. Comments provided in these discussions lead Stats SA to amalgamate several CPI releases pertaining to different geographical areas (e.g. rural and urban), while also introducing a more compact release known as "CPI at a glance". Of more relevance from the perspective of microdata access, were discussions on the statistical release associated with the annual General Household Survey (GHS). In discussions with participants, it emerged that there were several regular users of GHS microdata, who had never even seen the GHS release.

7. Stats SA also monitors visits to and downloads of documents and data from its web-site on a monthly basis. The use of Stats SA's website has grown dramatically over the past few years, while there has been a simultaneous drop in telephonic requests for statistical information. However, not all users desire electronic products. Surveys of

subscribers to hardcopy releases, particularly compendia of statistics such as the annual "Statistics in Brief", have shown that approximately half of these subscribers still prefer to receive statistical releases and reports in hard copy.

8. A common refrain from users is for more geographically disaggregated data – with respect to both social and economic data. This is particularly true of officials in provincial and local government, and is not surprising given pressures felt most acutely at local level for improvements in service delivery by government. Stats SA's ability to respond to these requests is limited by resources which in turn limit the size of sample surveys, as well as protecting the confidentiality of survey respondents.

9. While the confidentiality of information provided by respondents is protected through the Statistics Act, Act 6 of 1999, studies have indicated that many South Africans are concerned that information they provide to Stats SA in the course of a census or survey may be disclosed to third parties. This emerged through research undertaken in preparation for South Africa's next housing and population census to be conducted in 2011, on the attitudes of respondents towards Stats SA in general and census-taking in particular. In both focus groups [2] and in the quantitative study which followed the focus group study [3] it emerged that there are respondents who do not believe that information given to Stats SA will be kept confidential. Focus group participants admitted that this has inhibited their truthfulness in providing responses in previous censuses.

10. Thus, in disseminating microdata, it is incumbent upon Stats SA not only to ensure that respondents cannot be identified, but also to ensure that perceptions around confidentiality are managed, or this could impact on the accuracy of data provided by respondents in surveys.

---

[2]See http://www.statssa.gov.za/census2011/documents/CensusPublicityResearch-FocusGroupsResults.doc

[3]See http://www.statssa.gov.za/census2011/documents/CensusPublicityResearch-SurveyResults.doc

11. An example of mediating between the need to protect the identity of respondents and that of meeting user needs is provided through Census 2006 data dissemination. Community Profiles were provided in SuperCross format down to sub-place level (21 219 areas across South Africa, often corresponding to suburbs and villages) [4]. Local government officials demonstrated how the inability to disaggregate further geographically restricted their ability to identify areas most in need of services and undertake spatial planning for the municipality, given inequalities in the level of services provided to different areas rooted in the country's history. It was felt that despite having released data at Enumeration Area (EA) level for the 1996 census, confidentiality might be compromised in repeating this for the 2001 census. Thus a set of just over 56 000 "small areas" containing at least 500 households were created from grouping the approximately 80 000 EAs used for data collection in Census 2001. A selection of pre-aggregated variables corresponding to these areas were made available in digital format. [5]

Providing access to microdata: examples

12. In this section, some examples of microdata products distributed by Stats SA are outlined. These encompass microdata associated with the 2001 Census, household sample surveys, an administrative source of data used in the compilation of vital statistics and price data as an example from the domain of economic statistics. As mentioned above, files containing microdata are in general too large to be downloaded via the web and are usually distributed via CD. Stats SA's pricing policy [6] means that these CDs are made available at no or nominal cost to users. When there are new microdata available, past users are alerted individually to their availability, although the data are available to all users.

Census 2001 unit records

13. The Census 2001 "10% sample of unit records" [7] was compiled from selecting approximately 10% of all households living in housing units as well as 10% of all

[4] See http://www.statssa.gov.za/census01/html/C2001CommProfile.asp
[5] See http://www.statssa.gov.za/census01/html/C2001smallareastats.asp
[6] http://www.statssa.gov.za/about_statssa/pricing_policy.pdf
[7] See http://www.statssa.gov.za/census01/html/C200110percent.asp

individuals living in collective living quarters. Household information, information on all persons associated with the 10% sample of households and those living in collective living quarters as well as mortality information for the households forming part of the sample are packaged along with information on imputations and other metadata. A complete record of the copies distributed is not available, as some copies were distributed freely at workshops, but it is estimated that around 50 copies have been made available. Despite the census information now being five to six years old, there were still 944 downloads of "Census in brief" in the year ending 1 July 2007. Electronic microdata appears to be a 'specialist' product within the South African context at present.

The General Household Survey (GHS):

14. The GHS is an annual survey of approximately 30 000 households, and results in an annual statistical release in which information is provided at national and provincial level. Confidentialised unit records pertaining both to households and individuals, along with household and person weights respectively, are distributed on CD. Since 2004 the sample design has enabled users to obtain estimates at the level of South Africa's 52 district municipalities and six metropolitan areas. To give an indication of the extent of use of the product, 46 copies of the CD containing the 2005 GHS microdata have been distributed to date. By way of comparison, the corresponding statistical release, which was released in May 2006, was downloaded 2 868 times from the Stats SA website between 1 July 2006 and 1 July 2007.

The Labour Force Survey (LFS):

15. Currently Stats SA's LFS is conducted biannually, also using a sample of approximately 30 000 households, resulting in two statistical releases per year which provide information on the labour market at national and provincial level. The two most recently released sets of confidentialised unit records are for the March 2006 and September 2006 LFSs; 42 and 24 datasets respectively have been distributed. The downloads of the corresponding statistical releases in the year ending 1 July 2007 number 5569 and 3480 respectively. (As an aside, the uptake of both the statistical release and the

microdata product demonstrate the relatively long 'tail' of user requests/downloads for the LFS publications.)

16. As there is an overlap in the dwelling units included in successive LFS data collections (each dwelling unit is visited for five consecutive surveys), there is a possibility of providing time series records. Stats SA recently undertook an experiment to link these records, and from this constructed a product called "The Labour Force Surveys Panel". The product was release initially as a beta product, with requests for evaluation from recipients. 13 copies of these microdata were distributed during June 2007.

Administrative records as microdata: causes of death:

17. Statistics on mortality and causes of death are compiled by Stats SA through the processing of Death Notification Forms (DNFs) through which deaths are registered. Confidentialised unit records are provided on CD, packaged along with metadata and CSPro to enable users to compile their own tables. The processing of DNFs for deaths occurring from 1997-2002 lead to the release of the first such product in 2005, and this was followed by the unit record data for deaths occurring during 2003 and 2004 which were released in 2006. The statistical release provides analysis largely in terms of the underlying cause of death, while the unit records include all causes of death (these may number up to five) recorded on the DNF, as well as the derived underlying cause of death.

18. The province in which the death occurred and the province in which deceased resided at the time of death were omitted from the unit records published in 2005 and 2006. However, there was strong lobbying by the users of these data within the health sector to include this level of geography in the unit records, as this was felt to be important in terms of informing the formulating of relevant health care policy. After careful consideration of the risk of possible identification of individuals, particularly for those deaths associated with the smallest province (in terms of population, i.e. the Northern Cape), it was decided to include this information in the unit records released in 2007. The argument that prevailed was that that a user would have to bring other sources of

information to bear on the unit records, if information contained in these records were to be linked to a deceased person.

An example from economic statistics: price change microdata

19. Stats SA's CPI statistical releases carry only the aggregated price changes for various product groupings. A request was made to Stats SA by a doctoral student at a South African university for individual records reflecting the change in price of a specific product as sold at a particular outlet. The research proposal was shared with Stats SA, and the motivation for the request for such data included an indication that there had been a precedent of such data being provided by a national statistical organization elsewhere for similar research. The research involved applying data mining techniques to large numbers of price change records, to identify triggers of price changes. Following the signing of a confidentiality agreement, a CD containing price change records was provided to the researcher. The records did not include the details of the retail outlet to which the price changes referred. An agreement was also made that the researcher would present the results of the analysis to Stats SA staff.

Future trends in dissemination of microdata?

20. The examples above demonstrate that while Stats SA does provide microdata, this currently forms a small proportion of the uptake of statistical information products by uses, if we measure uptake in terms of *number* of users (and assume roughly one user per digital dataset and the same per publication download). Is the ratio of users of microdata to those who primarily rely on a document containing explanatory text, selected graphs and tables likely to grow?

21. Microdata users tend to be concentrated in the research community, with the tools and capacity to undertake their own statistical analyses. A rough analysis of the subscribers to time series data, as well as the subscribers to Stats SA's weekly electronic newsletter, shows that the research and education sector constitute a little less than a third of these time series users and newsletter subscribers. Anecdotal evidence from the national

statistical organizations of more developed countries suggests that students and researchers are the main users of the statistics they produce. Perhaps this user segment will grow with time in South Africa, and this will lead to greater interest in the use of microdata.

22. On the other hand, one of the reasons for requesting microdata is to produce cross-tabulations not included in standard statistical products. With increasing availability of products that enable users to create the tables they want, exactly as they want them, without direct access to the microdata, some users may no longer require large sets of unit records. Stats SA makes use of Nesstar, PXWeb and SuperWeb to enable users to create tables, graphs and maps, as well as cross tabulate data. [8] Stats SA will be monitoring the use of these tools via its website and will also continue to monitor requests for data packaged with tools such as SuperCross.

23. Metadata is important for the meaningful analysis of statistical data in general, and especially for microdata. Stats SA has embarked on a project (the Data Management and Information Delivery project, or DMID) to put in place policies, standards and tools to improve the quality of its statistical data outputs. The standardization and availability of metadata is a key element of the project. A tool supporting the capture of structured, standardized metadata pertaining to a series or psurvey instance has been developed [9]. The metadata which currently accompanies microdata will in time be standardized and structured, enabling the discovery of statistical data by users which meets their requirements, and facilitating the interpretation of the metadata and hence the associated statistical data.

24. Trends in the dissemination of microdata will certainly be informed by developments in technology. At least equal in importance is dialogue with users and potential users of statistical information products and services, to guide producers of statistics to meet user

[8] http://www.statssa.gov.za/timeseriesdata/main_timeseriesdata.asp
[9] See http://www.unece.org/stats/documents/ece/ces/ge.40/2007/wp.14.e.pdf for the current status of the DMID project.

requirements. Stats SA exists because of the Statistics Act (Act 6 of 1999)[10] . The purpose of the act is "to advance the planning, production, analysis, documentation, storage, dissemination *and use* of official and other statistics…". As a national statistical organization, we are obliged to ensure that statistics produced are also used; this means providing statistics to users in formats that are readily usable by them.

[10] http://www.info.gov.za/gazette/acts/1999/a6-99.htm