WP.39
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iv): Panel discussion on microdata protection versus remote access facilities

PANEL DISCUSSION

MICRODATA PROTECTION VERSUS REMOTE ACCESS FACILITIES

Chair: Jane Longhurst, Office for National Statistics, United Kingdom

One of the main functions of National Statistical Institutes (NSIs) is to publish detailed information on a large spectrum of aspects of the society. For this they collect large amounts of data via questionnaires and by using other registers or administrative sources leading to very rich databases. These databases are the main source for compiling their statistical publications.

The classical outputs are sets of marginal tables, published in volumes accessible by all users. However, these large databases can serve a second very important need. The rich databases are ideal material for performing statistical research. The results of these research projects give new insights and contribute to the well-being of modern societies.

Therefore it is a natural task for the NSIs to facilitate this research in the most effective way by providing access to these microdata files. But on the other hand the NSIs have a strong commitment to safeguard the privacy of the individual respondents in their databases. This commitment is not only an obligation from a Statistical Act in many countries, but also part of the ethical code for statisticians. An important side effect is also to guarantee the cooperation of the respondents in the future. This is why NSIs take confidentiality issues very seriously.

Within this discussion session specific focus will be given to three different approaches to managing the disclosure risk of microdata releases; licensing, masking techniques and remote access. Each approach has pros and cons and impacts on the microdata detail and therefore subsequent analysis in different ways. Panelists will provide an overview of their experience and thoughts on microdata access and participate in discussion from the floor. The aim will be to determine future directions on this issue and establish where research should be concentrated.

In order to stimulate discussion the thoughts and position of each panel member are provided below.

Paul Jackson, Office for National Statistics, Segensworth Road, Fareham, UK,
paul.j.jackson@ons.gov.uk

- Decisions about modalities for research data access should **begin** with consideration of the purpose for providing access.

  UK experience (and we are not alone) is that the demand for a sophisticated quantitative evidence base for policy making is increasing dramatically. Pre-defined standard tabular outputs can not meet this demand.

  Interdisciplinary (cross-cutting) and longitudinal analysis is needed to describe complex phenomena such as 'social capital' or 'child development and well-being'. Micro-data is essential to this research.

  Regional policy requires a tailored response to local issues. A central NSI can not hope to meet the specific demands of local evidence-based policy making.

  Analysis for policymaking is not founded on one-off reports, but is continuous.

- Decisions about modalities should then **acknowledge** the increasing competence and capacity in data management and analysis outside the NSI.

  Typically, academia now rivals or exceeds the statistical analysis competence of NSIs.

  Concepts for social and economic analysis (such as social capital) often emerge from non-government sources.

  Data management standards, organisational and technical, are now excellent in most universities and research institutions.

  The people of the analytical professions move much more freely now between NSIs and academia as their employers, taking their competences with them and spreading their skills amongst their peers.

- Decisions should then **recognise** that:

  Central, regional and local government will source analysis from the place best able to provide it - there is no closed shop for analysis.

  NSIs can not meet their wider objectives without enabling others to join in the description of the economy, society, and environment.

  NSIs have the privilege of lawful authority for original data collection. This privilege should benefit as many researchers as possible.

- These considerations lead us to **challenge some orthodoxes** for research data access methods:

  Fully anonymising micro-data before access will prevent good cross-cutting analysis and data linking.

  Removing local geographic identifiers will prevent sub-regional analysis.

  Sophisticated analysis can not be done through the letter-box of remote job submission.

  Access needs to be continuous for analysis supporting on-going policy development and evaluation.

  The people outside the NSI, after screening, can be as safe as the people inside the NSI, as are their facilities.

  The volume of research needed by public policy making means that checking every output can exhaust the resources of the NSI.

- **UK is working through this analysis and we have some conclusions:**

  Our new law concerns itself with the 'fitness and properness' of the researcher. A fit and proper researcher can have access to any data necessary for the described project.

  We will require researchers to check their own outputs (but we will provide standards to follow, and a support service)

  We will operate a three-tier model of facilities - data archive, academic laboratory/approved off-site environment, and on-site laboratory.

  We see little demand for remote job submission and have no plans to develop such a facility in the foreseeable future.

  We do see a role for remote access technologies in separating access from geography

Luisa Franconi, [3] Istat, DCMT, Via Cesare Balbo 16, 00145 Roma, Italy, e-mail franconi@istat.it

The need to maintain a diversity of types of access: different users, different needs, different access

Masked microdata sets obtained through the application of different protection methods offer an alternative to tabular outputs. These could take the form of Public Use File (PUF) available for all users or Microdata File for Research (MFR) dedicated to scientific research. Besides these products, recently many agencies are offering services to allow access to microdata sets, obtained by suppressing all direct identifiers in the original microdata, for scientific research purposes. These services take the form of Data Analysis Centres where only the output of the analysis of each researcher (and not the input microdata) is reviewed for confidentiality checks. Finally, some agencies have started offering also remote access or remote executions services for scientific research purposes to their survey microdata, sometimes applying some disclosure techniques. Although this latter type of access is extremely appealing for both researchers and statistical agencies this could not substitute the microdata products mentioned before for different reasons. First, there is a reason stemming from the right of access. All citizens have the right of access to information so, if from one side it is important to give priority to research, from the other we cannot forget other categories of users who may need microdata. Among them there are students and the global project of teaching statistics to train next generations to make further use of statistical data in more and more aspects of society. There are also analysts, marketing experts or even lawyers who may not appeal to the scientific research goal but for whom the data are important. Therefore it is crucial for the society as a whole to allow different channels of microdata access not to discriminate different types of research or different needs of the citizens. Secondly, remote access requires, from the point of view of the agencies, dedicated staff with experience in reviewing output of researchers. Two problems may arise here. Staff is costly and, in times of strict financial laws for public administration, it is hard to ask agencies to invest on this sector although crucial for research. So, if the initial burden of setting up such systems of remote access can be taken by the statistical agencies, in the long term is absolutely necessary to gain additional forms of financing in order to achieve sustainability in management. Additionally, the staff we are looking for managing such systems is well trained and with experience. This is demanding for many reasons: first there aren't still available general rules and guidelines for output checking consequently there is lack of automatic ways of checking complex output. These experiences still need to be built. This is a long process and problems here need to be addressed. Resources are needed also in the creation of masked microdata files but, in this case, it is more a one case necessity rather than a continuous process. Therefore, although appealing, remote access should not be the only way to give access to microdata.

Future research: data utility and user needs

Among the several item in the research agenda I see two main issues that, in my opinion needs, to be tackled. The first issue is improving data utility as one dimension of quality. We are releasing nowadays microdata sets that wouldn't have been released few years ago. This is because there has been a big research effort in defining what was more essential for the statistical agencies, preserving confidentiality, in order to increase the quantity of microdata offered to users. Measures of risks have been defined and procedure to assess them have been implemented, protection methods developed. Now it is high time to seriously improve further the quality of the products we are offering. One essential dimension of quality in the area we are dealing with is data utility as is it a duty of the data provider to give practical, relevant and useful measures of goodness of the released microdata to guide the users in measuring reliability of their analysis. Thorough investigation of the effects of protection methods is still lacking and definitions of methods that preserve essential statistics should be fostered. This will allow to put into practice the right balance between risk and utility in microdata release. The second issue is trying to address better user needs. On example of a need that is emerging in Europe is the level of the geographical details in released microdata. More and more studies want to investigate phenomenon at very detailed geographical level to have a clear map of the different characteristics of small areas and how this areas influence social and economical behaviour. Most surveys though cannot possibly be significant at such level. As territorial information are extremely identifying protection methods that still produce useful data while protecting confidentiality need to be further investigated together with clear guidelines on the use of such information and their statistical errors. Another clear user need is improving access to different types of data such as enterprise microdata as well as panel and longitudinal microdata.

Addressing these needs and finding ways to allow better access to such data is certainly the current challenge.

Anco Hundepool, Statistics Netherlands, Division of Methodology and Quality, P.O. Box 4000, 2270 JM Voorburg, The Netherlands, Email: ahnl@cbs.nl

History

When PCs became available and statistical analysis of micro data was possible for researchers, the need for access to micro data grew. In the beginning the NSIs were very reluctant to give access to these individual databases because of confidentiality reasons, but gradually methods have been developed to assess the disclosure risk and also to avoid disclosure. Methods like global recoding, local suppression and several perturbative methods. Many of these methods have been implemented in μ-ARGUS.

Special databases have been developed for researchers, who, after signing heavy contracts and under special conditions, could analyse these research datafiles on their own computers. These projects have been very successful, but on the other hand the need for more detailed datafiles remained.

The next step was to open Research Data Centres (RDCs), also called OnSite laboratories, etc. Different names, but the principle was the same. Researchers were given access to the rich less protected datafiles on computers within the premises of the NSIs. Without any possibilities to bring any material outside of the centre, the researchers could analyse these datafiles. The drawback is that any output they want to take home, has to be checked by the NSI-staff for disclosure risk. This is a heavy burden, but on the other hand it has made many research projects possible. Another drawback is that researchers have to travel many times to the premises of the NSI and the NSIs have to reserve expensive locations for these RDCs.

As modern information technology progresses, it is now possible to build secure connections over the internet, enabling Remote Access to Statistical Information. This has led to investigations to see whether these connections could be used as an alternative for the RDCs. The first prototypes for this have been built and the results look promising. Time is now ready to further investigate these possibilities and use it on a much wider scale. Giving access to databases to researchers in other countries or even to European databases now becomes within reach.

Conclusions

Modern researchers require easy access to the statistical databases for their research, while the NSIs have to preserve the confidentiality aspects of the respondents. Producing confidentialised microdata files for research has been the answer to these research needs in the eighties and nineties. Also safe research data centres, within the premises of the NSIs, have been an answer, but modern information technology makes it possible to build safe OnLine Access via the internet.

This approach is becoming very popular in a short period of time. The researcher can work from his own institute without travelling, while the NSI can still control the confidentiality of the output, as all results are still checked for confidentiality by the NSI. The production of protected micro data files will be obsolete very quickly, as all researchers will go for the Remote Access approach!

References

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Longhurst, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf (2006), *CENEX handbook on Statistical Disclosure Control,* CENEX-SDC project, http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handbook.pdf

Wolf, Peter-Paul and Anco Hundepool (2007), *Remote Access (not) at Statistics Netherlands*, ISI-session, Lisbon