DGINS 2005/91/2.1/EN

91ST DGINS CONFERENCE

COPENHAGEN, 26 + 27 MAY 2005

Disseminating statistics to the Research Community

Pedro DÍAZ MUÑOZ
Director of Statistical methodologies and tools, Eurostat, European Commission

Disseminating statistics to the Research Community

Pedro DÍAZ MUÑOZ
Director of Statistical methodologies and tools, Eurostat, European Commission

The dissemination of statistical data to researchers is influenced by several tensions. The risk of identification of individuals, the enormous information contents of the data not fully used, and the perception of the society both as data reporters and data users are the most relevant factors that, being in contradiction, entail these tensions. After an analysis of these factors, the paper discusses the measures taken at the EU to facilitate the access to data by researchers and proposes a set of actions for future improvement.

1. The information contents of statistical data

1.1 The gap between what is published and what could be published

Statistical data contains much more information than what a statistical office normally publishes. Consider first the case of a sample survey. In principle, all possible cross tabulations of the categorised variables collected can be produced. This means that the number of possible aggregated data cells can be obtained by the product of all the possible occurrences of the categorised dimensions. There are two types of limitations for a release of data with such a detail. On one hand, the size of the sampling errors prevents to estimate indicators on too small domains as their accuracy will be low and the values thus misleading. On the other hand, the risk of identification of the individuals is higher as the number of statistical units in the sample diminishes. Thus allowing secondary use of the microdata faces two different risks. That of identification of the individuals and that of erroneous inferences due to the lack of accuracy of the information produced.

The second case is that of data obtained from an exhaustive investigation such as a census or a register. Again the possibilities of producing aggregated data can be enormous and can be quantified as the number of cells in all possible cross tabulations of the categorised variables collected. Theoretically no accuracy problems are to be expected here even though there might be some response errors that will have an impact mainly in those cells which have very few individuals represented. The identification problems however are even higher than in the previous situation since any matches found refer with certitude to single individuals.

Another important information feature of statistical data is their potential by means of the linking of different data sources at micro level through the use of common identifying variables. An additional risk should be added in this case to those already mentioned above. This risk derives from the hypothesis underlying the data matching and the eventual lack of coherence of the data sources. It is worth noting that in certain cases some of the data sources used for the linkage have not been created for statistical purposes and consequently their quality may not be fit for the purpose of the statistical investigation.

The above-mentioned cases show that the amount of information that can be extracted from statistical data is much larger than what a statistical office normally publishes through all the dissemination means. Provided that precautions are taken in order to prevent the risks already mentioned, it is a responsibility of statistical offices to facilitate the secondary use of the data in order to optimise its utility.

1.2 Examples of secondary use of statistical information.

There is fortunately a wide diversity of uses of statistical data. Thus, trying to be exhaustive in describing them is an impossible task. I will just give some examples.

The proceedings of the 19th CEIES Seminar: Innovative Solutions in providing access to microdata [1], Lisbon on 26-27 September 2002 provide several examples of use of statistical data for research purposes. Three of them can be specially mentioned. First, Richard Blundell who shows how the use of individual longitudinal data on health, ageing and retirement can help to analyse the potential effects of population ageing and the economic impact of increasing demand of health and social services as the ageing phenomenon progresses. The data used in this case corresponds to the English Longitudinal Survey on Ageing (ELSA).

Second, Robert Erikson, provides an interesting example of cross-referencing three data sources to obtain a distribution of average marks among pupils from the working class who finished their compulsory education and the probabilities of transition to upper secondary school. The sources used were data from the 1990 population census on parents' occupation, average marks from a dataset of school leavers, and information on higher secondary education from a file listing admissions to this education level.

Third, Niels Westergaard-Nielsen stresses the use of linked employer-employee (EE) data which makes it possible to study differences across firms and the reactions to various policy interventions. The linked data answers policy-relevant questions on where jobs are created and where destroyed; and on how policy interventions affect firms and their demand for labour. The paper presents a very interesting description of the requirements of the data sources for the EE linkage to be possible.

Furthermore, the table below presents a synthesis of the projects reported by those research institutions which, during 2004, submitted to Eurostat requests of micro-data of the European Household panel (ECHP).

Research contracts using ECHP data. Year 2004. Main Topics	
Studies of specific sub-populations	*Studies of specific phenomena*
• Elderly • Poor • Regions • Long-term unemployed • Married women • Female participation in labour • Divorced • Temporary Workers • Persons at end of working life • Youth	• Mobility • Income inequality • Transition employment <-> unemployment • Taxation, subsidies • Intra-family transfers • Inequality in income and education • Wage changes • Education and Health • Labour market participation and fertility • Childcare • Discrimination

A very interesting example of use of micro data is EUROMOD, this is a tax-benefit model based on household micro data. This model estimates the effect of changes in social and fiscal policies on measures of personal income and household welfare. It has been used to evaluate European and National policies. For more information: http://www.econ.cam.ac.uk/dae/mu/emod.htm.

The examples shown above, should not give the impression that researchers are solely interested in microdata. There is an important class of researchers whose main interest is to analyse the economic evolution and to identify relations between the macro-magnitudes. Their work needs extensive amounts of historical series in raw form, that is, without any treatment such as seasonal or trading day adjustments which could perturb the basic signal to be extracted.

1.3 Abuses of statistical data

Fortunately not many examples can be found of bad use of statistical data. Regarding confidentiality breach, the reason could be that the protective measures taken by statistical organisations are sufficient. It could also be that statistical information doesn’t permit easily to carry out identifications in a sufficient scale that would make this a profitable practice.

Nevertheless, it is worth describing some possible fictional bad use of statistical information as this could be very damaging to the trust of society on the statistical system and, consequently, all means must be put for their prevention. This is done in the next paragraphs; note that the scenarios presented go beyond disclosure breach issues to others linked to misleading information or incorrect treatment of data.

- Listing of persons with high income levels can be extracted from a register of individuals including economic information, such as tax registers. These registers could be further linked with other statistical datasets such as census samples or budget surveys.

- Journalist attack against well known people (on the principle of the right to information) (for instance, “Some newspapers in Finland do take the trouble to process (and publish in the net) data on the richest persons (usually the names of the 1 000 persons who had the highest incomes and exact data on their yearly earnings, capital income and property value); local newspapers[1] have published all available tax data concerning all persons in their local area. The same income variables are also used in the Finnish EU-SILC data”).

- Nosy neighbour scenario: try to identify his neighbour in a micro file on the basis of a few key variables (type of dwelling, number of persons, sex, age, occupation, …).

- Linking micro or very detailed economic information to registers of individuals in order to provide a file for marketing purposes.

- Obtaining ratios of economic information of enterprises in order to identify sectors on which specific examination of individual companies can be done for administrative purposes.

- Misinterpretation of statistical results by persons not aware of basic statistical principles (for instance, deduction of causal relation where there is spurious correlation).

- Inconsistency between statistics derived from perturbed released micro data and official aggregated statistics.

- Production of aggregated data based on too few observations thus rendering it meaningless.

[1] These newspapers have been summoned to appear on court accused of acting against the law on personal data protection!

1.4 Protecting the Confidentiality

There is an extensive scientific literature that describes methods for protecting the data and analysing its disclosure risk. I refer to the work session on statistical data confidentiality held in Luxembourg on 7-9 April 2003 [2] in which most of these methods were presented. It is also worth noting the CASC project under the research Framework Programme 5 that has recently been completed [3]. The following table gives a schematic overview of the different methods currently used.

Methods to protect confidentiality of microdata files	
• Global Recoding	Several categories of a variable are collapsed into one.
• Local suppression	One or more values in an unsafe combination are replaced by a missing value.
• Top and bottom coding	Global recoding in case of ordinal categories.
• Post randomisation	Deliberate misclassification by changing the value of one or more categorised variables.
• Microaggregation	Replacement of individual quantitative values with values computed on small aggregates.
• Noise aggregation	Adding random noise to quantitative data.
• Data swapping	Exchange of some variables between two registers that have common categories for some predefined variables.

2. Supplying data to researchers

2.1. The trade-offs

The present section sums up what has been said so far. Whenever supplying data to researchers several trade-offs have to be taken into account:

(1) The identification risk of the data and its sensitivity. One can distinguish here two types of identification risks:

- the nosy neighbour scenario mentioned above: knowing few characteristics of one reporting unit, there is an attempt to identify it and release this information.

- the register attack: a register-commercial database is accessible. This allows identifying reporting unit (name, address) on the basis of few key variables like gender, activity, region, etc. The matching of the register and the micro file could allow for identification of a significant number of statistical data providers.

Nowadays, the second type of risk is considered the most serious because of the current technological developments that permit that a parallel database is developed for non-statistical purpose.

Protection from confidentiality breach is ensured by legal acts that contain provision for disclosure control. Nevertheless these acts provide general rules. Their practical interpretation varies very much from country to country and results in a high diversity of levels of protection.

(2) Capacity of analysis of information. As mentioned in Section 1.1 above, statistical organisations are well aware that what they publish is just small portion of all the information contained of the data. There is a big responsibility of the statistical organisation to facilitate secondary use of this rich information in order to meet specific needs.

(3) Perception of privacy of the society. Beyond the actual risk of identification is the risk perceived by the society. While the legal provisions for disclosure control may be sufficient or even excessive, the perception of the reporting units of the risk of ill use of the data they provide is very high and this could deteriorate the quality of their responses as they may try in the future to hide or distort some characteristics that they wouldn't like to be identified.

(4) Interest of the society in the information. In many cases detailed statistical analysis related to small populations or small areas has a very high interest for policy purposes. As in many other instances a contradiction may occur here: While individuals are concerned about the privacy of their responses, they are at the same time unhappy about this information not being fully exploited in order to identify societal needs and address imbalances in the distribution of wealth or public services.

2.2 Access to microdata for scientific purposes in the European Union. The Regulation 831/2002

In order to meet the needs of researchers in the EU, two instruments have been developed in the frame of the basic confidentiality legal acts (Regulations 1588/90 and 322/97). These two instruments are on one hand the statistical confidentiality committee that has the implementation powers in all confidentiality matters and the Commission Regulation 831/2002 concerning access to confidential data for scientific purposes. The reader can find a detailed description and analysis of this legal act in the paper presented by John King and Jean Louis Mercy in the Work Session on Statistical Data Confidentiality held in Luxembourg on 7-9 April 2003. While this regulation sets important hopes for the availability of microdata to the research community, its implementation has faced several difficulties which have made its development progress at a slow pace.

The statistical confidentiality committee of December 2004 has analysed the progress in the implementation of this Regulation and has agreed on the development of quick procedures to process the requests of researchers and to grant the eligibility of research institutions. At present microdata for researchers can only be provided for two statistical domains. These are the European Community Household Panel (ECHP) and, since very recently, the Labour Force Survey (LFS). In addition, the Community Innovation Survey Working Group is now discussing criteria to distribute microdata files of this investigation. Furthermore, a task force has been set up to do the same exercise for the coming Survey on Income and Living Conditions (EU-SILC). In parallel, resources have been allocated so that the backlog in treatment of requests is reduced and the processes are followed with the objective of having an improvement of the situation in the short term.

3. The international reflection. The UN/EC Task Force on Confidentiality and Microdata

In June 2003 the Conference of European Statisticians created the Task Force on Confidentiality and Microdata which was chaired by Dennis Trewin of the Australian Bureau of Statistics. The planned outputs of this activity were the development of agreed principles on the provision of access to microdata and the presentation of case studies of good practice consistent with those principles.

In May 2004 the Task Force produced a discussion document [4] which addressed the perspectives of NSIs and researchers and how the tensions between these perspectives could be resolved and discussed the different means such as anonymised microdata files, remote access facilities and data laboratories which could be used for that purpose. A set of principles were proposed and some issues for discussion were brought forward.

In October 2004 the Task Force compiled the comments by countries on a paper [5], providing a summary of these comments. It also included a listing of those issues in which there was a broad agreement and those for which a range of opinions was expressed.

The Task Force plans to end its mandate presenting a document of guidelines of good practice to the CES plenary of June 2005.

4. A Way Forward

Several paths can be taken in parallel in order to make the statistical data more useful for researchers while providing sufficient guarantee of non-disclosure. I will develop them now:

(1) Harmonised criteria for disclosure risk. In general the legislation at national and European levels is fairly harmonised with respect to what is considered confidential data. However, when applying this legislation, the criteria used differ considerably from country to country. These criteria have sometimes an important historical weight; sometimes do not have a solid scientific basis; and in many cases lead to conservative solutions because real risks are not well mastered.

This diversity of interpretations is a consequence of the fact that there is no harmonised approach of disclosure risk. To agree on disclosure risk, one should agree first on the sensitivity of the data (how "private" are the variables in the file) and on the possibility to match these data with external sources, that is, to the presence of key variables or identifying variables. Second, there is a need to find a harmonised way to measure the risk. Methodological work is needed to reconcile the different approaches or to express preference for one of them.

It is obvious here the need to have common core criteria which, while providing a satisfactory harmonisation level, allow for a degree of flexibility to adapt to the specific perception of the society in each country. This will also have the advantage of having a more solid internationally agreed basis that better justifies national choices made in the release of microdata.

(2) The eligibility of researchers and research projects. At present the criteria for eligibility both of the research project and of the researchers are not clear and of course not homogeneous throughout the European Union.

In assessing the researcher, one often tends to assess the research body to which he belongs and try to make a strong link between the researcher and the research body which is responsible for the former. A priori eligibility assessment is one step that deserves much care but an ex-post assessment based on a standard follow-up of the institution and the keeping of records about this institution seems to be more promising from an administrative and qualitative point of view. The possibility of having a black list including those that have ill-used the data would help to keep pressure on research institutions.

In assessing the research project, the involvement of the technical units which might already have some contact with the researchers and their work, which might have already conducted studies on the topics or might have a direct interest in the study is essential. Note that in some countries such as the UK, the aspect of interest for the statistical office is prevalent. Other aspects such as the originality of the research, the real need for confidential data, the absence of conflicting interest between NSO and the researcher can be taken into account.

Under the frame of Regulation 831/2002, criteria for eligibility have been developed and the corresponding evaluation questionnaires have been designed. These could form the basis of a set of transparent criteria that will ensure equal treatment of the scientific community throughout the EU.

(3) Legal provisions in case of ill-use. An important aspect of the protection of the data lies in the awareness of the user of the legal responsibilities that he incurs and on the legal actions that can actually be taken in the event of ill-use. In the case of international use, legal responsibilities have to be established and explicitly communicated.

(4) The role of the CENEX on statistical disclosure control. The task force on Centres of Excellence set up by the SPC has proposed to launch during 2005 a pilot project on the concept of Centres of Excellence (CENEX). Briefly, this concept consists of setting up a team of national statistical organisations that will provide expertise on a specific domain, developing tools or knowledge that will benefit not only the team members but the rest of the ESS community. Statistical disclosure control has been considered one of the two subjects that will integrate this pilot phase of CENEX. Eurostat is at present preparing the documentation to launch this project. The results expected at the end of the pilot, which is planned to lapse one year, will be an inventory of needs, the corresponding development of computer tools and a handbook of common practices for disclosure control.

(5) The code of practice that has been recently approved by the Statistical Programme Committee includes some provisions about the use of a statistical data by researchers and the protection of confidentiality. This code of practice will provide a framework to develop an harmonised activity in this domain.

(6) Remote access for researchers. There is a broad agreement among countries that this is a very promising approach. Nevertheless, the current experiences are rather isolated. One can consider two types of remote access. A first type can be considered remote execution: the researcher submits a programme and the output is later sent to him by email. A second type is properly remote access: the researcher performs the analysis and can immediately see the answer on the screen. Eurostat presented to the IT Directors' Group of October 2004 an analysis of three current experiences of remote access. Two of them, from the US National Center of Health Research and

from the Luxembourg Income Study project concern remote execution, while a third one, from Denmark Statistics is an example of remote access.

(7) Public use files. The question of whether it is possible to protect the confidentiality of a file to the point that the risk of breach can be considered sufficiently low so that public access to the anonymised data is possible has been very frequently asked. In the European Union there are examples of public use files in countries such as the UK, France, Italy and Austria. Of course, public use files do not replace confidential files, the latter giving much more possibilities of the analysis. On the other hand, public use files are produced at a minimal cost and the supply to users can be immediate.

Eurostat presented to the Statistical Confidentiality Committee in December 2004 a strategy for the creation of public use files and a specific proposal related to the Labour Force Survey (LFS). However, this proposal didn't meet with sufficient support from the Committee.

5. Conclusions

(1) There is an important gap between the information contained in statistical data and what a statistical office actually releases. A way to fill this gap is to supply microdata files to researchers.

(2) There are also many risks that have to be mastered - these are related to the legal protection of identification of individuals; to the possibility of ill use of the data; and to the perception of individuals of abusive manipulations of their information. These risks should be well managed.

(3) The objective is to fill the gap as long as the risks are satisfactorily managed. For this purpose several legal and technical measures can be explored. The legal measures concern eligibility of researchers and research projects. The technical ones refer to the different methods of confidentiality protection.

(4) International reflections show that although there is a broad consensus in favour of this supply of microdata to researchers, there is a diversity of views on many of the more detailed issues. In particular, criteria for considering a file sufficiently safe for dissemination vary widely.

(5) The EU Regulation 831/2002 is a useful legal frame for the supply of microdata to researchers. After a difficult initial implementation period, the process and delay established by this act will be examined in order to consider possible ideas for improvement.

(6) Several lines can be explored to improve the dissemination of microdata to researchers. First, the development of harmonised criteria for anonymisation and for eligibility of

researchers/research. Second, the legal frame to prevent ill use. Third, the application of the code of practice. Fourth, exploring the possibility of remote access to microdata. Last, the creation of public use files.

References:

[1] (2003), 19th CEIES seminar: Innovative solutions in providing access to microdata, Lisbon, 26 and 27 September 2002, Luxembourg: Office for Official Publications of the European Communities

[2] (2004), Monographs of official statistics: Work session on statistical data confidentiality, Luxembourg, 7 to 9 April 2003, Luxembourg: Office for Official Publications of the European Communities

[3] Domingo-Ferrer, J. & Torra, V. (2004), Privacy in Statistical Databases – CASC Project Final Conference, PSD 2004, Barcelona, Catalonia, Spain, June 2004, Proceedings, Springer-Verlag

[4] Trewin, D. (2004), CES Task Force on Confidentiality and Microdata – Discussion Paper, 52nd Plenary Session, Paris, 8-10 June 2004, UN Economic and Social Council, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians

[5] Trewin, D. (2004), CES Task Force on Confidentiality and Microdata, First meeting of the 2004/2005 Bureau, Washington, DC, 18-19 October 2004, Statistical Commission and Economic Commission for Europe, Conference of European Statisticians, Item 10