Business Microdata Dissemination at Istat

Ichim, D., Franconi L.,
Istituto Nazionale di Statistica
Piazza dell'Indipendenza, 4
00185 Rome, Italy
{ichim, franconi}@istat.it

The mission of National Statistical Institutes (NSI) is to produce reliable, impartial, transparent, accessible and pertinent information. The dissemination of this information should be performed in full compliance with the legislation pertaining to the privacy and confidentiality of respondents.

Nowadays the demand of analysis of microdata is steadily increasing. The Italian National Statistical Institute (Istat) adopts different dissemination channels in order to satisfy the users' needs. In this paper we will analyse several aspects of the dissemination of microdata files. Istat disseminates both restricted or unrestricted microdata files. The formers are usually designed to be used for scientific research, while the latter are public use files, freely available on the web.

National Statistical Institutes, the main producers of statistical information, have considerable experience in disseminating restricted-use microdata stemming from social surveys. Besides particular legal provisions, the main obstacle in releasing microdata files stemming from business surveys derives from the increased difficulty in protecting the confidentiality of respondents, i.e. businesses and enterprises. Indeed, classical disclosure limitation methods are ineffective when applied at high intensity, as is typically necessary for business microdata. Perturbation-based strategies or synthetic data (see for example Drechsler 2012) are usually more suitable for variables characterized by extremely skewed distributions.

At a first glance, the dissemination of detailed information and the preservation of the confidentiality of respondents might seem two very conflicting objectives. Anyway, by a careful analysis of the product to be released, the right balance may be found. In this paper we'll present the approach used in Istat for business microdata dissemination. The three main parts of the statistical disclosure control (SDC) process, risk assessment, disclosure limitation method and the quality assessment, are illustrated.

Finally, two important harmonisation dimensions are explored. Firstly, the release of multiple microdata files from the same survey is discussed. In Istat, a project is running aiming at the dissemination of both restricted microdata files for research purposes and un-restricted public use microdata files stemming from the same survey. Secondly, the harmonisation of microdata releases at European level is addressed. The model used by Istat proposes to achieve harmonisation of SDC process through the harmonisation of methodology at the input and the provision of harmonised and objective measures for the output allowing *de facto* some flexibility to the whole process and allowing for better adaptation to national context and better global efficiency.

1. Business microdata access at Istat and the policy of coherent information

The Italian legislation allows access to both social and business microdata. The data collected by Istat may be accessed at microdata level (individual data) using different channels in full compliance with the regulations pertaining to the privacy of respondents. Istat access portfolio includes the Research Data Centre, microdata files for research purposes and public use microdata files. In this section we

briefly describe these access channels; more details may be found on the Istat web-site, i.e. http://www.istat.it/en/products/microdata-files.

1.1 Network of points of access to the Research Data Centre

In 1999 Istat created the Laboratory for Analysis of Microdata (Laboratorio ADELE, an abbreviation of Analisi Dati ELEmentari), an on-site facility in Rome where researchers perform statistical analyses on original confidential microdata files stemming from social and business Istat surveys. The most popular statistical software packages (Stata, SPSS, SAS and R) are available and the access is free of charge. Access to the Laboratory is controlled and supervised and the final output of the research is released after checking for confidentiality by Istat staff. The results of the research cannot be considered official statistics. There is no limitation on the type of analysis that can be carried out at the Laboratory; this allows for more in-depth analysis of phenomena being studied, especially as far as business data are concerned. The most analysed business microdata stems from Structural Business survey and Communitiy Innovation Survey (CIS). Since 2012, Istat has increased the offer by creating a network of points of access to the Research Data Centre in each Istat regional office (currently eighteen), thus diminishing the travelling and subsistence costs related to any project, and by developing integrated business microdata (such as the Linked Employer Employee Database) to be analysed by users. The possibility of adopting the RDC in RDC approach with IAB is currently under study (see Bender et al. 2013).

1.2 Microdata Files for Research Purposes

The first microdata files for research (MFR) purposes were released at the beginning of 2009; two economic surveys, CIS and Structure of Earnings Survey (SES) were considered. Since then, besides several social microdata files, different MFR stemming from business surveys were released; they include Continuous Vocational Training Survey (CVTS), Farm Structure Survey (FSS) and Factors of Business Success (FOBS).

Access to ADELE or MFR can be granted to researchers working for universities or research institutions or fellows of bodies to which the "Code of conduct and professional practice applying to processing of personal data for statistical and scientific purposes ".

1.3 Public Use Files

In April 2013, Istat released for the first time four public use files, called mIcro.STAT, freely downloadable from Istat web-site. CIS was again included among these four files. The release of mIcro.STAT perfectly suits the Istat policy of developing a system of microdata that allows for access to coherent information. This means that, for each survey, the MFR is derived from the dataset available at ADELE, while the mIcro.STAT is derived from the corresponding MFR.

Istat adopts an unique production process for the release of both MFR and mIcro.STAT, the latter being public use files, freely downloadable from the Istat web-site. The existence of such process implies important efficiency gains from several points of view: a) the disclosure risk, b) coherence of the informative content of the files to be released, c) the physical creation of the microdata files and d) coherence of the associated metadata. The mIcro.STAT public use files are derived from the corresponding MFR by subsampling (Casciano et al., 2011). Both disclosure risk and some data utility requirements are taken into account when determining the optimal allocation. The second step of the mIcro.STAT procedure consists in drawing a random balanced sample, thus aiming at the approximate preservation of some weighted totals. Such Istat procedure aims at the release of mIcro.STAT files that maintain some quality indicators and, simultaneously, at the preservation of the advantages of dealing with random samples: the mIcro.STAT microdata are representative for the entire population, as the MFR does. Secondly, coherence with already published information is assured. For example, the equality between the published totals, the totals derived from the MFR and those derived from the mIcro.STAT are guaranteed. Just to mention two obvious advantages of this restriction, this latter

quality condition/indicator increases the trust in the dissemination strategies. Moreover, this coherence between estimates contributes to disabling disclosure scenarios based on differencing. Instead of totals, some other descriptive statistics might be used as well. Only published totals are dealt with by Istat since they are one of the most important statistical products and they are among the first statistics to be computed, an example for business survey is discussed in Foschi et al. (2012).

1.4 Istat SDCpolicy: data utility, confidentiality and comparability

Ideally, the results of statistical analyses performed on the original and disseminated microdata files should be the same. In practice this is not always possible due to the presence of highly visible enterprises (singled out during the risk assessment phase of the process) that share a high disclosure risk and that have to be treated via the application of disclosure limitation methods (see Hundepool et al,, 2012). Nonetheless, with respect to a specific survey, users' needs are taken into consideration by conducting a rigorous review of the previous data analyses. This classification of data analyses highlights the data characteristics to be (exactly) preserved during the disclosure limitation process. Different statistical disclosure control (SDC) techniques are applied in order to manage the disclosure risk of each statistical unit. For the MFR dissemination at Istat, the choice of the SDC techniques is simultaneously guided by the disclosure scenarios and data utility. In Table 1, a summary of the protection techniques used for the above mentioned business microdata MFR is given. In order to preserve as much as possible the analytical validity of the business microdata files, only the units considered at risk of disclosure are generally modified.

Table 1: summary of statistical disclosure limitation methods applied to enterprise survey microdata

Survey\SDC	Rounding	Individual ranking	Recoding	Perturbation
CIS	X	X	X	X
SES		X	X	X
CVTS	X		X	X
FSS	X	X	X	X
FOBS	X		X	

Finally, a special type of users' needs is comparability. Some surveys are harmonised at European level in order to ensure comparability of statistics among different Member States and over time. The users would greatly benefit from a harmonised dissemination of microdata files, too. A possible strategy to achieve this harmonised dissemination is the selection of a set of benchmark statistics or quality indicators, as proposed in Ichim and Franconi (2010) and discussed in Franconi and Ichim (2012). The comparability benchmark statistics could be (among) the statistics to be maintained, as derived from analyses of the research potential of microdata.

2. Benefits and problems of accessing microdata

National Statistical Institutes (NSIs) disseminate many statistics stemming from a single survey. In the information technology era, the number of access channels developed and used by the NSIs rapidly grows. To cite only few of them, Istat, commonly uses a) internet-related channels (e.g. dynamic graphics, storytelling, RSS feeds, Facebook and Twitter, electronic newsletters), b) main media channels (e.g. TV, radio, press), and c) mobile communication technologies (e.g. special apps). One drawback of all these channels is that they provide the same information, only the format and the provided related services may vary. In this section, by related services we mean, for example, the documentation and its level of detail or the possibility to directly contact Istat staff for (quick) questions. Of course, when statistics are disseminated using TV channels, the documentation is represented in "breaking news" formats; when using Twitter, Istat staff will twit to the users' comments; when using Istat web-site, an e-mail address is commonly provided in order to allow users to send their comments/observations, etc. Anyway, the information provided is the same. Only the

degree of complexity of its presentation may vary. On the one hand, this strategy guarantees that the disseminated information is coherent though the many dissemination channels. On the other hand, dissemination and communication of "value added" products requires making choices about what data or results to highlight and on which findings and implications to focus on, which statistical methodologies to apply and so on. These choices performed by the NSIs imply several consequences. First of all, only part of the collected information is returned to the society. For example, for the press release corresponding to the Community Innovation Survey (CIS), 2008-2010, about 75-80%[1] of the registered variables were used to derive the reported statistics. Secondly, the neutrality may not always be guaranteed. For example, there are many situations where the best methodology is not well-defined. In such situations, the NSI has to make a choice, anyway. Moreover, the choice of the information to be published as aggregated statistics may be guided by subjective criteria, even if reliability and degree of significance of statistics are two commonly used criteria. Thirdly, transparency may be questioned. Indeed, when using dissemination channels with well-defined and constraining formats, there is no always enough "space" to provide detailed information on the missing data treatment, choice of the methodology, choice of the auxiliary information, etc. Anyway, due to the quantity of information collected by the NSIs, it is probably impossible to disseminate statistics on each variable collected using the most detailed documentation. Moreover, well-structured, well-designed and coherent aggregated statistics are generally sufficient to satisfy the information needs of the general public.

Nevertheless, there are categories of people that indicate more often their need to access more detailed information. Among these categories, it is worth mentioning scientific researchers, decision-makers and enterprises/business. Access to microdata allows researchers to develop in-depth analyses aimed to better understandings of the economic environment and the complex mechanisms governing economic development. Indeed, economic analyses are generally based on models of individual behaviour specifying the objectives and constraints faced by the economic agents. Additionally, by using microdata, researchers may choose which analyses to perform, which subsets of data to consider, which methods to apply, which variables to derive/construct and may apply whatever multivariate analysis.

By continuing with the CIS 2008-2010 example, at the beginning of the press communicate it is stated that about 31% of the Italian enterprises successfully introduced innovation in the 2008-2010 period. If the data were analysed by economic activity, it could be observed that there is a large variability, as shown on Figure 1 – Overall. This analysis could be possible if detailed aggregated statistics were published in the press communicate. It could be then deduced that 100% of the NACE 5 (extraction of minerals) enterprises are innovating enterprises. This is an example of misleading aggregated statistics information since it is based on a single sampled unit. Obviously, the reliability of the published statistics may be assessed only using microdata. A different strength of microdata may be observed by analysing in-depth the types of innovation. Indeed, the distribution of the type of innovation does not seem to be equally distributed among the economic sectors. In Figure 2, we illustrate these distributions for three NACE divisions, namely 10, 12 and 32. The simplest observation is that service and logistics innovations are completely absent for the enterprises in Nace 12. Again, this kind of basic analysis is possible only when, for each enterprise, the information on the introduced type of innovation, if any, is available. This level of detail of information may be found and used only if access to microdata is guaranteed. In absence of access to microdata, statistical analysis for research purposes may be misleading.

[1] Mastrostefano, V. (2013), Personal communication. It should be mentioned that through Eurostat web-site, all aggregated statistics are available.

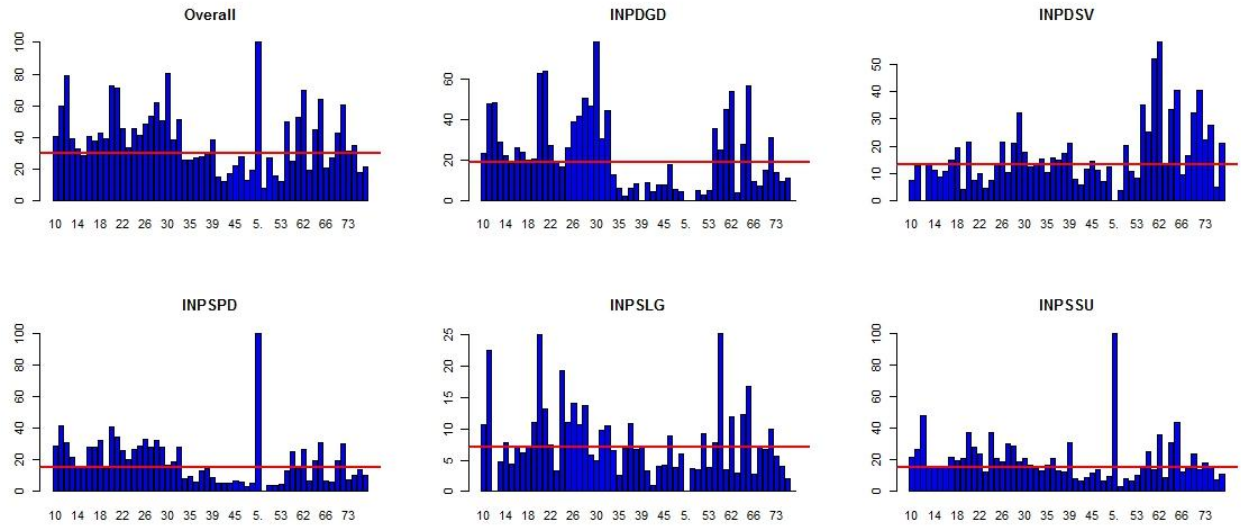

Figure 1. Percentage of innovating enterprises in the 2008-2010 period, by different definitions of "innovating enterprises". Overall = enterprises introducing some innovation, INPDGD= product innovation, INPDSV = service innovation, INPSPD = production process innovation, INPSLG = logistics innovation, INPSSU = other innovation.

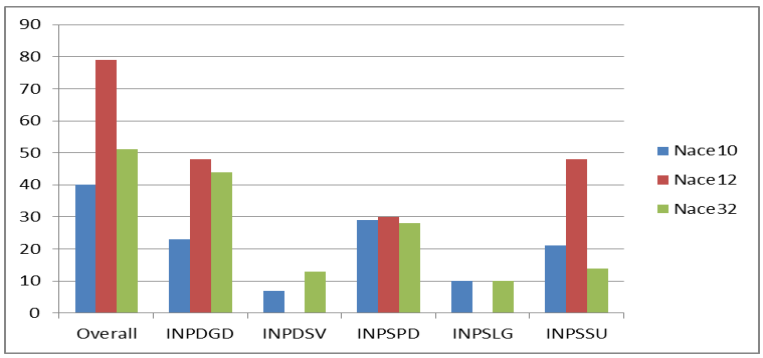

Figure 2. Comparison of distribution of the type of innovation for three NACE divisions.

It is sometimes argued that enterprises do not access the NSI statistics because small and medium-sized businesses would generally like to have a few current statistics focused on their own activity. By default, microdata products could represent a perfect answer to this need since they contain much more detail than the aggregate statistics. For example, the most detailed information contained in the CIS 2008-2010 press communicate refers to less than 30 Nace divisions, while the microdata files MFR and mIcro.STAT contains about 50 Nace divisions (subsequent to recoding). The microdata accessed at the ADELE represents the original, pure, microdata file. At Istat the release of business microdata is seen as a way of returning relevant information to the enterprises as the access to microdata is governed by the same protocol for both private and public entities. The only obstacle is represented by the fact that microdata files for research may be accessed only by entities\departments legally recognized as developing research activities. Two main solutions are generally adopted by enterprises to overcome this obstacle: a) cooperation with academics and, more recently, b) the usage of public use files which may be downloaded from the Istat web-site since April 2013.

It should be stressed that the correct usage of microdata depends on two main factors: documentation and knowledge and skills. With respect to the former, Istat microdata are provided together with a system of metadata including methodological notes referring to both survey and statistical disclosure process, classifications, examples of structure of the files, questionnaires and some basic software codes to allow an initial data exploration. Much more could be done in terms of data presentation, which indeed lacks of friendliness. Istat runs and coordinates a project for the creation of the Italian Data Archive (Istat, 2013) aiming at improving several aspects of the microdata management. These

aspects are related to data presentation, documentation, data discovery, and data usability. We are also aware that knowledge and skills of many business users in order to autonomously and correctly use NSI statistics may represent an obstacle. Istat is continuously participating to many initiatives related to statistical literacy and different projects aiming at transforming data into statistical commodities.

Other problems of microdata dissemination are surely represented by the costs associated to the production and dissemination of microdata or to running a research data centre. Staff, experience, knowledge and infrastructure are only few aspects that should be taken into account. A National Statistical Data Archive could surely improve the general management of microdata dissemination and could very much help data users, by allowing them to quickly find the high quality official statistics data. It is generally considered that transparency, impartiality and neutrality can be increased by providing regular access to microdata files. The benefits obtained by the research community and, in medium-long term, by the society are not under discussion.

3. Conclusions

In this paper we briefly illustrate Istat system of access to business microdata. Each component of Istat system, i.e. the Research Data Centre, the microdata files for research purposes and the public use files, is briefly described. Istat disclosure policy for the access to microdata files stemming from economic surveys is introduced as well. We also provide a quick overview of benefits and problems of accessing microdata files. It is our opinion that microdata products may represent a valuable product to be considered by both scientific research community and enterprises.

References

1) D. Ichim, L. Franconi, Achieving Strategies to Achieve SDC Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys, Privacy in Statistical Databases Lecture Notes in Computer Science Volume 6344, 2010, pp 284-296, DOI 10.1007/978-3-642-15838-4_25 Print ISBN 978-3-642-15837-7, Springer Berlin Heidelberg
2) L. Franconi, D. Ichim, Achieving Comparability of Earnings, Privacy in Statistical Databases Lecture Notes in Computer Science Volume 7556, 2012, pp 188-199, DOI 10.1007/978-3-642-33627-0_15, Print ISBN 978-3-642-33626-3, Springer Berlin Heidelberg.
3) C. Casciano, D. Ichim, L. Corallo, Sampling as a way to reduce risk and create a Public Use File maintaining weighted totals, Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spain, 26-28 October 2011, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011.
4) F. Foschi, M.C. Casciano, L. Franconi, D. Ichim, Designing Multiple Releases from the Small and Medium Enterprises Survey, Privacy in Statistical Databases, Lecture Notes in Computer Science Volume 7556, 2012, pp 200-215, DOI 10.1007/978-3-642-33627-0_16 Print ISBN 978-3-642-33626-3, Springer Berlin Heidelberg.
5) Istat, Micro-data: a crucial asset for statistical system, UNECE, Conference of European Statistician, Sixty-first Plenary Session, Challanges in providing access to microdata for research purposes, Geneva, 10-12 June 2013. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2013/31.pdf
6) Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P.-P. (2012) Statistical Disclosure Control, John Wiley & Sons, Ltd, Chichester, UK.
7) Bender S. and Heining J., Franconi L. and Ichim, D. Microdata access: an international perspective, Deliverable 7.2 Blue-ETS project 2013, available at http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable7.2.pdf

8) Drechsler, J. New Data Dissemination Approaches in Old Europe - Synthetic Datasets for a German Establishment Survey. Journal of Applied Statistics, 2012, Vol. 39, 243–265.