Tabled Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (iii): Modes of access to microdata

Istat experience on releasing multiple microdata files stemming from the same survey

Prepared by Daniela Ichim and Luisa Franconi

Istat experience on releasing multiple microdata files stemming from the same survey

Daniela Ichim and Luisa Franconi

[*]Istat, Piazza dell'Indipendenza 4, 00185, Rome, Italy, e-mail: ichim@istat.it, franconi@istat.it

Abstract: Since April 2013, the Italian National Institute of Statistics has enriched its microdata release system with a new product, mIcro.STAT, a public use microdata file available for download at Istat web-site, http://www.istat.it/en/products/microdata-files#file_microstat. This paper analyses Istat microdata access system from different points of view - design and production process, characteristics and metadata – describes the experiences gathered so far and presents the challenges ahead.

1 Introduction

Official statistics aim to satisfy the information requirements of different categories of users, which are continuously increasing and diversifying. National Statistical Offices are expected to develop integrated systems of dissemination that permit to combine data from multiple sources and increase the completeness of statistical information. The development of communication technologies and, more importantly, the increased need for information, lead to more users requiring access to microdata. These users differ greatly in their information needs and in their ability to manage statistical information. In this context, several challenges are faced by NSOs: they need to produce an increasing amount of consistent and relevant statistics; more timely data are needed on more topics than before; more detailed territorial statistics are more often demanded; integration of data from multiple sources is a necessity for assessing cross-sectional issues; finally new strategies and dissemination tools need to be developed to effectively supply information to different users.

The Italian National Institute of Statistics (Istat) adopts different dissemination channels in order to satisfy the users' needs. In this paper we'll present the approach used by Istat for microdata dissemination. 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.

2 Istat dissemination system

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.

2.1 The network of access points 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 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. Over the years, more than 280 research projects were conducted using the Adele infrastructure; about 20 projects per year involving data stemming from about 50 surveys. The most analysed microdata stem from Community Innovation Survey (CIS), University Graduates Vocational Integration, Labour Force Survey (LFS), Small and Medium Enterprises Survey (SME) and Population Census. 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).

2.2 MFR: Microdata Files for Research Purposes

The approach adopted in Istat for microdata dissemination is represented by the three main parts of the statistical disclosure control (SDC) process: risk assessment, disclosure limitation method and quality assessment.

When designing any microdata release we take into account both the research potential of the survey and Istat dissemination policy, see, for example Franconi and Ichim (2012). The risk assessment phase generally includes external register and spontaneous identification scenarios. 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 rare cases or highly visible units (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. Table 1, columns 3rd – 8th, shows a summary of the protection techniques used for the MFR dissemination. In order to preserve as much as possible the analytical validity of business microdata, only the units considered at risk of disclosure are generally modified. Finally, the quality assessment phase is carried out by the comparing statistical indicators computed using the original and anonymised microdata file.

The first microdata files for research (MFR) purposes were released at the beginning of 2009. Since then, MFR stemming from 13 business and social surveys were released. In Table 1, we provide the full list of surveys from which a file of microdata was released by Istat for research purposes. It should be mentioned that different subsequent waves may be available from a single survey. In such cases, the same anonymisation methodology is applied to subsequent survey waves, unless there are significant changes in the survey structure or in the number of units at risk of disclosure.

The second column of Table 1 indicates whether the same MFR is released at European level by Eurostat. In such situations, the approach described in Ichim and Franconi (2010) is generally applied. Besides the files listed in Table 1, Istat participates in the release at European level of additional microdata files: European Community Household Panel (ECHP), European Union Statistics on Income and Living Conditions (EU-SILC), Adult Education Survey (AES) and European Road Freight Transport Survey (ERFT).

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 ". With regards to a single survey, we notice that the number of projects conducted at ADELE does not necessarily decrease once an MFR is released. For example, as it can be observed in Table 2, the number of projects which required access to CIS or LFS data at ADELE remained almost constant over the years, while the number of MFR requests steadily increases. It may be deduced that once an MFR is released, the target audience increases, even if it is constrained by the fulfilment of very restrictive access conditions. Moreover, the projects for which the analysis of highly detailed data, i.e. projects requiring access to ADELE, seems to remain constant over the years. In could be concluded that MFR and ADELE are used by different category of users, even though the two services are designed for the same scientific research community.

Survey	EU	Variable suppression	Recoding	Top-bottom coding	Model-based perturbation	Individual ranking	Rounding
Continuous Vocational Training Survey (CVTS)	YES	x	x	x	x		x
Factors of Business Success (Fobs)	NO	x	x	x			x
Farm StructureSurvey (FSS)	NO	x	x		x	x	x
Graduates' Transition	NO	x	x	x			x
Labour Force Survey - cross-sectional quarterly (LFS)	YES	x	x				
Labour Force Survey-12 months longitudinal data	NO	x	x				
Population Census 2001	NO	x	x				
Road Accidents Resulting in Death or Injury	NO	x	x	x			x
Structure of Earnings Survey (SES)	YES	x	x		x	x	
Survey on Doctorate Holders' Vocational Integration	NO	x	x	x			
Italian Innovation Survey (CIS)	YES	x	x		x	x	x
University Graduates Census	NO	x	x	x			
University Graduates' Vocational Integration	NO	x	x	x			x

Table 1 Surveys for whom Istat disseminates microdata files for research purposes and adopted statistical disclosure control methodologies.

Year/Survey	CIS MFR	CIS ADELE	LFS MFR	LFS ADELE
2009	1	0	2	1
2010	1	2	4	4
2011	4	4	10	2
2012	5	4	34	0
2013 (till July)	3	0	19	3

Table 2 Number of projects requiring access to MFR and ADELE for the most analysed surveys, i.e. CIS and LFS.

2.3 mIcro.STAT: 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. Since then, other two public use microdata files were released. The list of these mIcro.STAT files together with the number of external[1] downloads is presented in Table 3. As it may be observed, a new category of users was reached.

Survey	**April**	**May**	**June**	**July**	**August**
Italian community innovation survey	125	118	71	43	30
Road accidents resulting in death or injury	89	65	29	34	22
University graduates census	90	46	23	12	7
Doctorate holders' vocational integration	51	24	16	8	3
University graduates' vocational integration			11	44	15
Graduates' transition				14	11
Total	**355**	**253**	**150**	**155**	**88**

Table 3 Number of downloads of mIcro.STAT by survey and month of download.

The release of mIcro.STAT is in line with 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 a unique production process for the release of both MFR and mIcro.STAT. 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. 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: mIcro.STAT

[1] Not performed by Istat staff.

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).

Istat public use files, mIcro.STAT, may be freely downloaded from the web-site http://www.istat.it/en/products/microdata-files#file_microstat. Users are required to register to Istat web-site and to provide a valid e-mail address. Before downloading the file, users are also required to accept three conditions of use: a) not to identify any statistical unit, b) to use the data only for statistical purposes and c) cite the source, as mentioned in the Creative Commons Licence. The microdata are provided in a tab-delimited format in a compressed zip file, together with the documentation needed to explore and analyse the data.

3 Documentation

The microdata files provided by Istat are enhanced by a documentation system. As Istat microdata system is under construction, several features of the documentation will probably change.

At the time of writing, the documentation provided together with MFR and micro.STAT includes a survey abstract, a survey methodology report, an anonymization report describing the changes due to the statistical disclosure control methods, the survey questionnaire, a layout of the file, classifications and tools for importing the data. The MFR documentation also includes an example of structure file.

The survey abstract briefly describes the main characteristics of the survey (phenomena, data collection, national and international regulation, etc.). The survey methodology report includes statistical details on the sampling design, data collection, data editing and calibration methods used to process the survey data. A bibliography is generally included, pointing to papers and web-sites where more information about the survey may be found. The anonymisation report includes a) details on the disclosure risk assessment (disclosure scenarios and units at risk of disclosure), b) detailed information on the statistical disclosure limitation methods applied; for each modified variable, a precise indication of the type of modification is provided, c) an analysis of the microdata file analytical validity, if needed, and d) a reference section listing papers and web-sites describing in more detail the anonymisation process. The anonymisation report also contains information on the number of records and the number of variables registered in the file and the sum of the weights. We included this information as a kind of control information to be checked by users when starting to analyse the data. The file layout and the classifications represent an important tool to explore the structure and the informational content of the microdata file. The file layout contains several columns describing, for each variable, its label, description, type (categorical, continuous), the format in which it was registered (numeric or alphanumeric), number of decimal, decimal separator, sign (+ or -) and measurement unit (if necessary). By clicking on the type of a categorical variable, a file describing its classification is opened. The file layout and the classifications may be opened by whatever browser. By means of the links between the file layout and classifications, the user may easily observe the detail of information contained in the microdata file. To further support the initial exploration of the microdata when restricted access in involved, an example of the structure of the file is freely available. Indeed, an example of structure file is a kind of toy example file automatically produced from the MFR and the file available at ADELE. The example of structure file contains a random number of records, no more than 100, the same variables as the corresponding restricted use file. In order to guarantee that no identification is possible, an unconstrained swapping is independently applied to each variable. Due to the swapping procedure and to the limited number of records, the example of structure file has no analytical validity. When sufficiently tested, routines for importing microdata are also provided to users.

The ADELE, MFR and mIcro.STAT microdata files documentation always contains the above described documents. The only exception is represented by the absence of the example of structure file from the mIcro.STAT documentation. The reasoning is that the mIcro.STAT is itself a freely available file.

We are confident that we have setup a documentation system that allows users to perform their analyses without any additional Istat support. Anyway, the documentation system currently has two main drawbacks. Firstly, except for the MFR files which are disseminated at European level (see Table 1), the documentation is available only in Italian. Secondly, as the documentation is provided by means of a compressed zip file, users do not have too much possibility to discover, search and explore the microdata files. We are aware that the translation of documents is a general issue for the European countries. As far as the data documentation is concerned, Istat is involved in different national and international activities aiming at improving the data documentation system. The main ideas of one such project is briefly described in the next section.

4 Improving access to microdata: new services to users

Microdata products are an essential part of any dissemination system. During these years of microdata dissemination at Istat, we observed that users need some additional tools and services for the analysis of data: the microdata files are almost useless without a proper context/framework which should be represented by the dissemination strategy of any NSI.

Defining a dissemination strategy has an influence on improving all stages of dissemination activities: characterising dissemination policies, designing products and services, preparing and presenting statistics, disseminating information on the website, and promoting and marketing products, services and “statistical releases” to users. To design a user-centric dissemination system for microdata, the governance, the infrastructure and the legal framework need to be defined and put in place together with practical data access policies and flexible microdata products and services (Istat, 2013). This also requires establishing different relationships with microdata users.

4.1 Governance

In Italy, the vision for a user-centric approach to microdata access is promoted by a group of institutions. These institutions are forming connected data clearinghouses and related infrastructure that will provide the broadest possible access to publicly funded microdata. This project is leaded by official statistics: Istat will be the hub of a joint venture that aims at developing a network of data archives that will deal with microdata produced using public funding. In this network the Bank of Italy and Istat will manage microdata for their respective topics, a third node will be in charge of research institutes microdata and the Ministry of Education will encourage universities to contribute their microdata. In the future, Istat should become the hub for microdata access for data from government departments (like ministries) and the Italian National Statistical System.

4.2 Infrastructure

In Italy several public institutions provide microdata access. Even though microdata are provided, they are dispersed across many government institutions or research institutes. In terms of infrastructure, the Italian network of data archives will develop a coherent and single access to the microdata products or services offered by public statistics. A web catalogue of all microdata accessible to users both directly through a download or via requests for further specialised services (remote execution, remote access, etc.) will be developed.

4.3 Legal framework

Istat, as a hub of the network, should have the possibility of providing access to microdata from other institutions of the national statistical system and managing remote access to microdata from research institutes. A legal framework supporting such architecture is crucial. This approach shall become possible in Italy following the amendment of the Italian statistical law being prepared at present.

4.4 Data discovery, accessibility

A user-centric system allows for a quick and clear overview of the kind of microdata that are available and under which conditions. User-friendly search protocols will be developed to help users finding the microdata they need.This will be made possible by the adoption and implementation of standardised metadata protocols, a milestone of any dissemination system. Finally, the network of data archives will work towards harmonisation of microdata access policies inside the Italian national statistical system to support the management of requests to different institutions.

4.5 Relationship with users

Besides the cooperation among the data archives, the focus of the network will be the collaboration with users and statistical literacy. The users of microdata need knowledge in statistics, analytics, data analysis and computer programming. The process of data analysis should be understood in its simplest form by most citizens. It should, therefore, be taught to students and should be the core competency for the staff of public administration.

Availability of powerful tools for analysing statistical microdata is not sufficient. Adequate competences should be developed for selecting suitable methods, applying them correctly and understanding and interpreting the results. Statistical literacy should be constantly promoted and supported by NSOs both inside their national statistical systems (government agencies, public administrations) and outside (undergraduates, master and PhD students, young researchers, etc.).

The network of data archives and Istat school of statistics (SAES) will collaborate to offer training programs network will pave the way for changes in the relationship between producers and users of official statistics. Partnerships are currently being created to collaborate with Istat, to improve survey design, to propose changes to questionnaires used in data collection and to improve usefulness of data. Users will not just be data analysts but will be increasingly called upon for an active contribution toward the improvements of official statistics.

5 Conclusions

In this paper we describe Istat integrated system of access to microdata and its future development. By illustrating the different access channels, including the latest product, mIcro.STAT public use file, we highlight the way Istat performs multiple releases from the same survey. Istat experience shows that different products are indeed used by different categories of users. In this paper we provided a brief overview of the number of statistical projects using each of Istat dissemination/access channels: ADELE, MFR and mIcro.STAT.

Data providers start being more active in improving the context in which the microdata is released. This involve an investment in metadata standards. Istat policy is based on the recognition that the industrialisation process will necessarily require the implementation of metadata standards for the whole life cycle. These same metadata may be re-used to offer advanced services to users: microdata discovery, search ability and exploration.

References

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

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.

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.

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.

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.

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

Istat (2013). Micro-data: a crucial asset for statistical systems. Economic Commission for Europe, Conference of European Statisticians, Sixty-first plenary session, Geneva, 10-12 June 2013.