Micro-Moments Dataset

linked micro-aggregated data on ICT usage, innovation
and economic performance in enterprises

NOTE TO THE RESEARCHER VISITING EUROSTAT'S SAFE CENTRE

1. INTRODUCTION

An authorised researcher has the right to use the Micro-Moments Dataset (MMD) for an agreed research project via the Safe Centre at Eurostat. This possibility depends on overall data availability at Eurostat, Member States' willingness to allow the linked and micro-aggregated output data to be offered for the research use and the permission for using the data for the particular research project.

Before entering to the Safe Centre the researcher (assisted by the respective Eurostat staff) need to ensure that she/he:

(a) before coming to Eurostat

□ has received and read the Eurostat Manual on the protection of confidential data;

□ has received and read the MMD description in *Bartelsman et al (2013) Cross-country analysis of ICT impact using firm-level data: Micro Moments Database and Research Infrastructure.*

□ has carefully read the Description of the Variables, Breakdowns and Coverage, available on the official MMD site

□ has signed the individual confidentiality declaration;

(b) once in Eurostat

□ has received service card and personal access card for the Safe Centre;

□ has been furnished by the user ID and a password

□ has been personally briefed by Eurostat's legal service and the Unit in charge of the linked output data (Unit G4 Innovation and information society)

2. EUROSTAT SAFE CENTRE

Eurostat Safe Centre visiting times are the standard office hours from 7H00 to 20H00.

Core office hours are 9H15 to 16H45.

During the core hours the assistance on PC, Eurostat's software, linked output data, etc. is guaranteed.

The beginning of the *first visiting day* of a researcher needs to be fixed in the core hours. Software procedures can be left running overnight although there are no guarantees on the problem free execution due to potential power cuts, network updates and so on.

The researcher works within a given directory of a stand-alone PC. The source data is stored under the directory which is dedicated only for the researcher and all the programs and the output will be stored in that directory. Duplicating of the source data sets or subsets of it shall be avoided in order not to extend unnecessarily the volume of the stored data.

Should there be another researcher coming for the stand alone PC before the work of the previous researcher is finished, all the work files (data, programs, interim results, etc.) will be moved away from the PC as it will be fully cleaned (reset) for the new user. When the first researcher continues the work the software needs to be re-installed and the work files need to be re-loaded. When working consecutive days or without interruption by another researcher (within short period) this will not be necessary. The working schedule of the researcher in the Safe Center needs to be agreed upon in advance based on the availability of the Safe Center in order to avoid disruption of researcher's work.

The available software is Microsoft office 2010, SAS 9.2 with SAS Enterprise Guide

4.3 and STATA 13. The SAS modules available are: Base SAS, SAS/STAT, SAS/GRAPH, SAS/ETS, SAS/IML. During the course of 2015 SAS will be upgraded to version 9.4 with Enterprise Guide 6.1. The STATA version is STATA/SE for large datasets.

In the Safe Centre it is **NOT** possible

- to print documents;

- to copy data to CD-ROM, DVD, Zip drives, USB keys or any other portable devices;

- to copy data to the local hard disk (in secure server environment);

- to connect recording devices to the serial, parallel and USB ports;

- to connect a laptop to the network;

- to use email;

- to make internet connections.

3. RESEARCHER'S OWN DATA AND OWN SOFTWARE

Should the researcher wish to use own data set(s), this will not happen without Eurostat’s explicit *prior* agreement. If the use of the researcher's own data set is indispensable for the project it is advised to contact Eurostat on that before delivering the application of the research project. Own data may concern for example background variables such as national totals or particular variables for separate economic activities.

Any data of the researcher needs to be uploaded onto the stand-alone PC from a CD-ROM or a USB-key by the Eurostat staff. The researcher her-/himself is in charge of the functionality of her/his own software. Under no condition can researcher copy any data from the stand-alone PC onto any information carrier device.

4. DATA DESCRIPTION

The MMD is a micro-aggregated output of several microdata-linking projects funded by Eurostat and implemented at the National Statistical Institutes (NSIs) of 14 European countries via an externally-coordinated coding procedure. From the national linked firm-level data sources, a cross-country industry dataset at a medium-level of aggregation (over a set of dimensions not available earlier) was constructed. This dataset includes measures of ICT usage and innovative activity together with measures of business performance and industry dynamics. These measures include typical aggregates, such as sums and means, but also higher moments of distributions of variables of interest, as well as moments from multivariate distributions.

The MMD variables were drawn/derived from 4 main data sources at the national micro-level:

- a set of variables pertaining to ICT usage by firms drawn from (1) the Survey on ICT Usage and e-Commerce in Enterprises

- a set of innovation variables coming from (2) the Community Innovation Survey

- a set of variables describing the economic characteristics and performance of firms drawn largely from the (3) Business Register and (4) Structural Business Survey.

For some countries a few economic performance variables were not found in a typical Business Register or Structural Business Survey and were linked-in from other sources: surveys in skills, international sourcing, ICT investment and innovation.

Variables that populate the MMD are described in detail in the Annex.

NSIs from the following countries participated in the data-linking initiative: Austria, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, the Netherlands, Norway, Poland, Slovenia, Sweden, the United Kingdom.

The years covered by the dataset vary among countries and are subject mainly to the availability of the CIS and Survey on ICT Usage and e-Commerce in Enterprises data. The longest period covers 2000-2010 (depending on the country and the variable). For more details, please see Final report on ESSLait Metadata Repository[1].

In order to get an idea of what the dataset looks like, please consult its aggregated version which is publicly available at the Information Society dedicated section of the Eurostat website (under the heading *ESSnet on Linking of Microdata to Analyse ICT Impact (ESSLait), 2013*): http://ec.europa.eu/eurostat/web/digital-economy-and-society/methodology. Please bear in mind that this version of the data does not contain all the breakdowns and only contains information over the 5 EU KLEMS-type groups of sectors: Electrical machinery, post and communication services; manufacturing (excluding electrical); Other production; Market services (excluding post and telecommunications), Non-market services.

The data sets are provided for the researcher's use in CSV format.

5. OUTPUT VALIDATION

The researcher should ensure that any results of the research published or otherwise disseminated do not contain information which may permit the identification of individual records of the data.

Any deliberate attempt to compromise the confidentiality of organizations to which confidential data for scientific purpose relate may result in prosecution in accordance with applicable law.

[1] Hagsten et al (2013), The Multifacetal Nature of ICT. Final Report of the ESSnet on Linking Microdata to Analyse ICT Impact, Eurstat, downloadable from the Eurostat official website http://ec.europa.eu/eurostat/web/digital-economy-and-society/methodology

All the results to be transmitted away from the Safe Centre are checked by Eurostat to avoid any disclosure of confidential data.

The non-identification covers both primary and secondary confidentiality.

Primary confidentiality concerns data, whose dissemination would allow disclosure of individual enterprise. The main reason for declaring MMD to be primary confidential is too few enterprises in a cell. Any statistics (tables, graphs, textual references) on any kind of sub-population (cell) shall not be published if they consist of less than 10 enterprises.

Secondary confidentiality concerns data which is not primary disclosive, but whose dissemination, when combined with other data permits the identification of an enterprise or the disclosure of an attribute of the enterprise.

Even if the confidentiality has been defined to be in the researcher's (or researcher's background institution's) responsibility, Eurostat validates all output the researcher wishes to export from the Safe Centre. The researcher shall be able to explain the processes and show that the output is non-disclosive. Data which has been validated is safe to use further outside the Safe Centre.

Eurostat checks also the consistency of the output produced in the Safe Center with the statistics made public on Eurostat website. General non-disclosure principle and the rules specified above shall be respected in all circumstances.

In the following there are some illustrative examples of primary and secondary confidentiality issues. Examples are drawn from another enterprise-based dataset of Eurostat, namely Structure of Earnings Survey data (SES). The examples and remarks below are given only as an indication and guideline for the researcher. They do not cover all the possible situations and possibilities.

Example of primary confidentiality

Primary confidentiality means that any cell of the output to be exported from the Safe Centre needs to fulfill directly the primary confidentiality condition above.

Table 1. Illustration of primary confidentially check

Region	Economic activity	Occupation	Median earnings	Number of enterprises	Number of persons employed
AA1	61	21	25.2	20	120
AA1	61	22	30.5	4	25
AA1	61	23	22.2	18	55
AA1	61	24	24.4	16	210
AA1	61	25	19.1	31	482
Total AA1	61	21-26	23.1	89	892

In the Illustration (i), the occupation 22 does not fulfill the condition that the published cell shall have 10 or more units in it. This output proposal would be rejected as the number of local units in occupation 22 is only 4.

Example of secondary confidentiality

Hiding the occupation 22 in the Illustration (i) would create a problem of secondary confidentiality: a reader would be able to calculate the number of the local units in the hidden cell using the total AA1 and non-hidden information. Also the hidden sensitive information, median earnings, could become easily estimated for the occupation 22, at least its range. This output proposal would be rejected even with hiding the information in the line of the occupation 22.

While protecting and validating the secondary confidentiality, the data in different *independent* tables and different forms of presentations (graphs) and *classification levels* and systems shall be taken into account.

Regressions and other forms of output

In general, regression results are non-disclosive at an exact level (some inferences may be drawn within a margin of error in particular cases). Moreover, this small risk can be reduced further in ways which do not significantly reduce the usefulness of the results. The simplest way is non-reporting of incidental parameters, such as estimated constants or the coefficients on irrelevant dummy variables. In general a regression with (N-K) → ∞ which does not report all significant parameters is non-disclosive for all practical purposes.

For other analytical results, the disclosive nature depends on the manipulations carried out. The assumption is that results are disclosive unless proved otherwise, and therefore it is in the researcher's interests to show that the results are non-disclosive.

Graphs and *quantiles* are treated as tables which just present the information in a different form. *Maximum* and *minimum* values are seen as tables with *x* enterprises in a cell i.e. they are confidential unless *x>10.*

Detecting and protecting secondary confidentiality for the other than tabular forms of output shall be ensured.

6. PrEPARING THE OUTPUT FOR VALIDATION

The proposed output shall include all the information which is needed for the output validation even if this (extra) information will not be published / used further (frequencies etc.). For example, the two last columns in the Illustration (i) are necessary for validating the data even if they would not be published in the final report the researcher will prepare. It is up to the researcher to decide the type of presentation and measures in showing that all the rules have been respected.

Together with the results to be validated, all the programs to derive this output will have to be presented as all the results must be reproducible.

It should be noticed that the output validation concerns only the disclosiveness of the data. Output validation is not a quality check. Appropriateness of the assumptions or underlying theory or analysis will not be assessed and remain the researcher's responsibility.

7. REJECTION OF THE OUTPUT

The proposed output will be automatically rejected if the confidentiality rules above are not respected. The output may also be rejected if it is not fully understood or the output is very long. In these cases Eurostat cannot be sure whether the confidentiality rules are fully respected and cannot therefore validate the data. Non-documented or unexplained output will not be approved (presentations of tables or other results alone).

Eurostat does not necessarily make proposals how to modify the output to get it accepted but just indicates the reason why the output has been rejected. If the output is rejected the researcher needs to re-work the output for having it re-validated.

Notice that the burden of proving that results are safe is onto the researcher. Further, it is in the researcher's own interest to show the output which can be validated within reasonable time and without potential costly re-visits to the Safe Centre.

8. VI OUTPUT VALIDATION FORM AND TIME

The output intended to be exported from the Safe Centre needs to be saved (together with the related programs) under given directory on the stand-alone computer in the Safe Centre. The table data shall be saved together with all the other data necessary for the validation with the adequate headings, titles and other metadata. New derived variables should be documented and meaningful variable names have to be used.

At the end of the research work, the researcher informs the contact person of Eurostat Unit G4 about her/his intention to export the data from the Safe Centre after which the validation process begins. For facilitating the validation work, it is recommended that the researcher in person shows the output to Eurostat before leaving the Safe Centre and explains its main characteristics.

Eurostat reserves the right to define the time needed for validating the output data. All technical and organizational measures will be taken by Eurostat to ensure an efficient checking without undue delays. Eurostat will make every effort to ensure that this delay does not exceed two weeks.

The validated output shall be sent to the researcher by email.

Read more on the Eurostat output checking procedure at:

https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf_en

References

Airaksinen et al (2008) Information Society: ICT Impact Assessment By Linking Data From Different Sources, Final Report, Eurostat, http://ec.europa.eu/eurostat/documents/341889/725524/2006-2008-ICT-IMPACTS-FINAL-REPORT-V2.pdf/72f0967d-a164-46ad-a6d0-246be5a6d418

Bartelsman, Eric J., John Haltiwanger, and Stefano Scarpetta, “Measuring and analyzing cross-country differences in firm dynamics, in Producer Dynamics, Timothy Dunne, J. Bradford Jensen, and Mark Roberts, eds., NBER Studies in Income and Wealth, Volume 68, University of Chicago Press, 2009, pp. 15-82

Bartelsman et al (2013) Cross-country analysis of ICT impact using firm-level data: Micro Moments Database and Research Infrastructure, Eurostat, http://www.scb.se/Grupp/OmSCB/Internationellt/Dokument/esslait-mmd-final.pdf

Brandt et al (2009) Guidelines for the checking of output based on microdata research, ESSNet Statistical Disclosure Control, http://www.cros-portal.eu/sites/default/files//GuidelinesForOutputChecking_Dec2009.pdf

Hagsten et al (2012), ESSnet on Linking of Microdata on ICT Usage, Final Report, Eurostat, http://ec.europa.eu/eurostat/documents/341889/725524/2010-2012-ICT-IMPACT-2012-Final-report.pdf/90cf5094-334a-4ff1-8f60-047c2d650c60