



**Microdata File
for Research**

Cummunity Innovation Survey (CIS4)

Years 2002-2004

Survey methodology

CONTENTS

1. Introduction	3
2. Target population.....	4
3. Sampling design	5
4. Data collection and data treatment	6
5. Weights calculation method.....	8
6. Dissemination of the survey results	9
7. Glossary	9
8. References	11
9. Contacts.....	12

1. Introduction

The Italian Innovation Survey (CIS) covers innovation activities of the Italian enterprises with at least ten employees operative in industry and services. In particular, CIS collects information on the type and novelty of innovation introduced, innovation outputs and inputs, as well as sources, effects, cooperation and hampering factors of innovation..

The Italian Innovation survey is part of the EU *Community Innovation Survey* (CIS), carried out on a two-yearly basis (four-yearly basis before 2004) by all EU Member States and candidate countries, plus Norway and Iceland. Since 2000, CIS has also become one of the major sources of data for the *European Innovation Scoreboard*.

Since 2004, CIS has been conducted under the *Commission Regulation No 1450/2004*, which establishes the legal basis for innovation statistics and makes it compulsory to deliver data every two years on some key variables.

In order to ensure a satisfying comparability across countries, the CIS is carried out on the basis of a standard core questionnaire and a harmonised survey methodology developed by Eurostat, in close cooperation with the participating countries. The methodological basis of CIS is provided by the OECD *Oslo Manual*.

In order to satisfy the users' needs, Istat disseminates microdata files for scientific research. The dissemination of the statistical information must be performed in full compliance with the regulations pertaining the privacy of respondents.

This document gives information on the survey methodology used for the Italian CIS4. In Section 2 the target population is described. In Section 3 the sampling design is illustrated. In Section 4 the CIS4 data collection and treatment methods are presented, while in Section 5 the methodology used for the weights computation is described. The last sections provide some information on the main channels for the data dissemination as well as the statistical definitions of innovation used in the survey and the key references for the survey development.

2. Target population

2.1 Economic activity - Nace

The following industries are included in the core target population of the CIS4:

- mining and quarrying (NACE 10-14);
- manufacturing (NACE 15-37);
- electricity, gas and water supply (NACE 40-41);
- wholesale trade (NACE 51);
- transport, storage and communication (NACE 60-64);
- financial intermediation (NACE 65-67);
- computer and related activities (NACE 72);
- architectural and engineering activities (NACE 74.2);
- technical testing and analysis (NACE 74.3).

“Non-core” industries that were covered in addition are:

- construction (NACE 45);
- motor trade (NACE 50);
- retail trade (NACE 52);
- hotels and restaurants (NACE 55);
- real estate activities (NACE 70);
- renting of machinery and equipment without an operator (NACE 71);
- research and development (NACE 73);
- other business activities: legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling; business and management consultancy; holdings (NACE 74.1); advertising (NACE 74.4); labour recruitment and provision of personnel (NACE 74.5); investigation and security activities (NACE 74.6); industrial cleaning (NACE 74.7); miscellaneous business activities n.e.c. (NACE 74.8).

2.2 Size-classes

All enterprises included in the target population follow the minimum coverage which is all enterprises with 10 employees or more.

The variable used for identifying size's classes is the average number of employees in the year 2004.

2.3 Statistical units

The main statistical unit for CIS 4 is the enterprise, as defined in the Council Regulation 696/1993 on statistical units or as defined in the national statistical business register.

3. Sampling design

The survey was based on a one stage stratified simple random sample. At least 3 enterprises in each stratum were selected. In the case of less than 3 enterprises in a stratum, a full census was conducted.

The target population of the CIS4 is broken down into similar structured subgroups or strata (which should be as homogeneous as possible and form mutually exclusive groups). The stratification variables to be used for the CIS4, i.e. the characteristics used to break down the sample into similarly structured groups, are:

1. the economic activities (in accordance with NACE). Stratification by NACE has been done at least at two-digit (division) level, except for NACE 24, 35, 74. NACE 24.4, 35.3, 74.2 and 74.3 groups were treated as separate NACE sectors while the remaining groups of NACE 24, 35 and 74 were treated as single NACE sectors;
2. enterprise size according to the number of employees. The size-classes used were the following: 10-49 employees; 50-249 employees; 250+ employees.;
3. regional aspects at NUTS 2 level. The regional allocation of the sample was taken into consideration when sampling. In particular, the breakdown of national territory into regions was performed on the basis of the NUTS level 2. (Regulation EC No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics - NUTS)

A multi-variable and multi-domain sample allocation was used. The adopted procedure was an application of the Bethel algorithm [5]. It was an optimum allocation since it aimed at minimizing survey costs under the constraint that sampling errors of estimates of each variable of interest didn't exceed the given upper bounds assigned to each of them. For the pursuit of the best allocation, three auxiliary variables were used: number of

employees, turnover and total innovation expenditure. In particular, previous CIS3 results were used, all referring to the year 2000¹.

To keep the response burden down, a coordinated selection technique (Jales sampling) was adopted in order to avoid the inclusion of the same enterprises in the sample over time.

3.1 Sampling frame

The official, up-to-date, statistical business register², called ASIA (Archivio Statistico delle Imprese Attive - statistical business register of active enterprises) was used. It provides both the key variables for the stratification (number of employees, NACE economic activity, NUTS geographical information) and the identification characters (enterprise name, address, etc.). With reference to the CIS4, the survey universe consisted of all the profit enterprises and independent professional units active at 31 December 2004.

3.2 Sample size and overall sample rate

The Italian CIS4 sample included 44,571 enterprises out of a population of about 193,300 enterprises with 10 employees or more and potentially active in the year 2004 and the average response rate turned out to be 49%.

4. Data collection and data treatment

Data were collected through a combination of census and sample survey. The census concerned all the enterprises with more than 249 employees: they were around 3,300 enterprises out of a target population of more than 193,300 units of target population.

The survey was conducted as a mail survey.

After having launched the survey in April 2005, a first reminder was sent out by mail to non respondents in September 2005. A second reminder has been carried out in November 2005 using a mixed approach: by phone (for the largest ones and those operating in sectors with a small number of enterprises) and mail.

Since non-respondents, as an unweighted percentage of all relevant enterprises in the sampling

¹ With regard to the strata not covered by the previous CIS (CIS3), such as the NACE division '45' (Construction), the auxiliary variables (employees and turnover) were taken from the Italian Statistical Business Register (ASIA).

² Council Regulation (EEC) N° 2186/1993 of 22 July 1993.

frame, exceed 30%, a simple random sample of 10% of the non-respondents was selected for a *Non Response Survey*, carried out through CATI interviews, on the basis of a short questionnaire, including some of the questions of the harmonised core questionnaire, in order to determine if the non-respondent is an innovator or not. Information from the *Non Response Survey* was then used when calculating the weighting factors.

During the collection stage, there has been a systematic and sustained follow up with the responding enterprises in order to obtain the missing information and to keep at a minimum items non-response. Several data quality checks both at micro- and macro-level were done before the results are finally processed. The checking procedures were developed in SAS. An ad-hoc software package was provided by Eurostat.

The types of the performed edits can be grouped into five categories:

- Completeness checks. These refer to the questionnaires that are not fully completed. Contact were made with the surveyed units to get the missing information.
- Out of scope checks concerning the units that do not belong to the target population (i.e. wrong NACE, wrong size etc). The out of scope units have been dropped from data processing.
- Domain or coding checks, aimed at verifying if the given answers were permissible (i.e. the answer is within the range of answers allowed). Such errors were corrected to bring it into line with the range allowed.
- Consistency checks to verify if there are some situations in which a combination to several questions is possible or if relationships between two variables are within specific bounds..
- Routing errors' checks concerning the right/wrong comprehension of the sequencing of questions.

The automatic editing procedures were carried out only after having solved all the missing values concerning the general variables on the enterprises and the key variables on innovation (introduction of product/process innovations). In case of the missing information in the first set of variables, Istat used information available in the business register or from other administrative or statistical sources, while for missing values in the second set Istat contacted the enterprises. Lastly, the checks and correction on R&D expenditures were done making use of information taken from R&D survey.

Before running the automatic corrections for the other items non-response (but only after having trying to get any missing data from the enterprise concerned), imputation was done. A first type of imputation was the logical (or deductive) imputation, that was applied to correct inconsistencies or for imputing some item non response, based on the answers provided in related questions. After having cleaned all the data, the SAS imputation programmes went on correcting metric variables separately from ordinal variables.

As far as the metric variables are concerned, a weighted mean of each metric variable, by NACE and size class, was calculated (just after having removed outliers) and applied as a ratio to the enterprises with the missing values within each stratum.

The imputation of the ordinal and nominal was done after the metric estimation. The technique used is nearest-neighbour hot decking. Nearest-neighbour imputation uses data from clean records (free of logical inconsistency) to impute missing values of recipient records. The donors were chosen in such a way that some measure of distance between the donor and the recipient is minimised. The distance measure is a multivariate measure based on the reported data.

After running the estimation procedure for nominal and ordinal variables, the next step was to run macro-level checks in order to check if the CIS4 results are in line with other published results. In particular, these checks were carried out by comparing CIS4 and SBS data, on one side, and CIS4 and CIS3 data, on the other.

5. Weights calculation method

Calibration estimators methodology [7], currently applied at Istat, was used for the estimation process. It can be applied to the extent that the known totals of some auxiliary variables, strictly correlated to the variables of interest, are available. These calibration estimators have the following properties: they are more efficient than the direct estimators because of the auxiliary information used; they reduce the bias effect due to the non response and under coverage; they produce estimates of auxiliary variables that equal the known totals of such variables.

The final weights are obtained by adopting the following procedure: an initial weight is assigned to each sampled unit with reference to the sampling plan as the reciprocal of the inclusion probability. Two correction factors for initial weights are then calculated: a first one is the unit non response factor; a second one is to satisfy equality between estimation

of auxiliary variables and known totals from the Register. The final weights are thus obtained as the result of the product between initial weights and correction factors.

For CIS, as well as for most of the business surveys, number of enterprises and number of employees were used as auxiliary variables, according to the information provided by the Italian Official Business Register ASIA.

The results of the non-response analysis have also been taken into account when calculating the final weighting factors applying the methodology adopted by Hansen e Hurwitz in order to reduce the non-response bias of the estimator concerning the number of innovating firms.

The software used was GENESEES, a generalised software implemented in SAS language by ISTAT researchers, available for all users from ISTAT website:

http://www.istat.it/strumenti/metodi/software/produzione_stime/genesees/index.html

6. Dissemination of the survey results

Besides the microdata file for scientific research, several statistical information products are available.

The Italian CIS4 aggregated indicators are disseminated mainly through the ISTAT website: see 1. www.istat.it and 2. http://www.istat.it/dati/catalogo/20080227_00/

The aggregated indicators of the European survey may be found at the New Chronos website. The researchers might analyse the non-perturbed microdata file at the Italian or European Research Data Centers. More information is available at the ISTAT and EUROSTAT web-sites: <http://www.istat.it/dati/microdati/adele.html> (e-mail: adele@istat.it) and <http://ec.europa.eu/eurostat>, respectively.

7. Glossary

The questionnaire, which is reproduced in file *ISTAT_MFR_CIS_Questionnaire_2004.pdf*, provides the definitions of many of the concepts specific to innovation. The main ones are repeated here, in so far as they are used to categorise enterprises as having innovation activity or not, and to determine different kinds of innovators.

Innovation

An innovation is a new or significantly improved product (good or service) introduced to the market or the introduction within an enterprise of a new or significantly improved process. Innovations are based on the results of new technological developments, new combinations of existing technology or the utilisation of other knowledge acquired by the enterprise.

Innovations may be developed by the innovating enterprise or by another enterprise; however, purely selling innovations wholly produced and developed by other enterprises is not included as an innovation activity. Innovations should be new to the enterprise concerned; for product innovations they do not necessarily have to be new to the market and for process innovations the enterprise does not necessarily have to be the first to have introduced the process.

Product innovation

A product innovation is a product (good or service) which is either new or significantly improved with respect to its fundamental characteristics, technical specifications, incorporated software or other immaterial components, intended uses, or user friendliness. Changes of a solely aesthetic nature are not included. Product innovations may have the following features:

- the new technology leads to a better performance of the good or service;
- a broadening of the product or service range is accomplished.

Examples are: change of materials in goods, introduction of ecological products, use of chipcard systems, electronic banking and insurance, web-related services and e-commerce (but only creating an information site without on-line services is not an innovation).

Process innovation

A process innovation includes new and significantly improved production technology, methods of supplying services and of delivering products. The outcome (of the process) should be significant with respect to the level of output, quality of products or costs of production and distribution. Purely organisational or managerial changes are not included. Examples are: order picking, tracking and tracing of shipments, connecting (data) communication and transport, bar-code systems, optical processing of data, expert systems, first use of CAD/CAE. ISO-certification is only an innovation when it is directly related to the introduction of new or improved processes, ordering systems, minimisation

systems for stocks, product interchange systems, transport-logistics, computer aided logistics.

Enterprises with innovation activity

Enterprises that have had any kind of innovation activity during the survey period, i.e. have introduced or implemented new products and/or processes, had abandoned innovation activity, or had on-going innovation activity at the end of the reference period.

Enterprises that introduced or implemented new products and/or processes during the reference period. A further classification of successful innovators can be made to distinguish between:

- those that had only introduced product innovations;
- those that had only implemented process innovations;
- those that had done both, introduced product innovations and implemented process innovations.

In all three of these sub-categories of successful innovators these enterprises may additionally have had on-going and/or abandoned innovation activity.

Enterprises without innovation activity

Those enterprises that had no innovation activity whatsoever during the reference period.

These enterprises only answered a limited set of questions from the survey in relation to: the absence of innovation activity; factors hampering innovation; patents and other protection methods; and other important strategic and organisational changes within the enterprise.

8. References

Bethel, J. (1989) "Sample allocation in multivariate surveys". *Survey Methodology*, 15: 47-57.

Deville, J.C. and Särndal, C.E. (1992) "Calibration estimators in survey sampling", *Journal of the American Statistical Association* 87, 367.382.

Eurostat (2004) "Innovation in Europe: Results for the EU, Iceland and Norway", Panorama of the European Union, Theme 9 Sciences and technologies, European Communities.

Eurostat (2006) “The Fourth Community Innovation Survey (CIS4). Methodological recommendations”, Doc. Eurostat/F4/STI/CIS/2b.

Hansen, M. H., and Hurwitz, W. N., (1946) “The Problem of Nonresponse in Sample Surveys”, *Journal of the American Statistical Association*, 41, 516-529.

ISTAT, (2008) “Statistiche sull’innovazione delle imprese”, Collana *Informazioni*, N. 1.

OECD (1997) “Oslo-Manual, Proposed Guidelines for Collecting and Interpreting Technological Innovation Data”, Organisation for Economic Co-Operation and Development, Paris.

Ohlsson, E., B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Kott e P. S. Colledge (eds.). “Coordination of samples using permanent random numbers”. In *Business Survey Methods*. Wiley, New York, 1995.

9. Contacts

For questions on the survey, please contact Valeria Mastrostefano, mastrost@istat.it.

For questions on the applied statistical disclosure control methodology survey, please contact Daniela Ichim, ichim@istat.it.

Authors

The document was written by Valeria Mastrostefano.