RESEARCH DATA CENTER
accesso ai microdati per la ricerca economica

Survey of Industrial and service Firms (INVIND)

Dataset handbook

Reference period for the data: 1984-2023

Updated in: July 2024

Keywords: Business surveys, sample microdata, industry and services

Compiled by: Sample Surveys Division of the Statistical Analysis Directorate

BANCA D'ITALIA
EUROSISTEMA

July 2024

Table of contents

General information

The Survey of Industrial and Service Firms ('Invind') is conducted annually on a sample of Italian firms comprising approximately 3,000 firms in industry excluding construction, 1,000 private non-financial service providers (the latter include trade, hotels and restaurants, transport and communication, real estate, IT and other private services) and 500 construction firms, to collect firms' records and structural data.

Firms are asked to provide mainly quantitative information on trends in key economic variables. The questionnaire covers employment, investment, turnover, profit or loss for the year, production capacity, debt. Furthermore, new focus sections on specific topics of interest are introduced each year.

Reference population

The reference universe consists of firms in industry excluding construction and in non-financial private services with 20 or more employees and of construction firms with 10 or more employees, with administrative headquarters in Italy.

Prior to 1998, the survey only covered manufacturing firms with 50 or more employees. Since 1999, the reference universe has been extended to include all industrial firms excluding construction and firms classified in subsection B (Mining and quarrying), D (Electricity, gas, steam and air-conditioning supply) and E (Water supply, sewerage, waste management and remediation) of the ATECO 2007 classification (derived from NACE Rev. 2).

In 2001, the survey was extended (with an abridged questionnaire) to include firms with 20 to 49 employees. In 2002, the reference population was expanded to include non-financial private service firms with 20 or more employees (while excluding from the market services category the following types of services: those provided by credit and insurance firms, public services and other social and personal services). In 2006, the survey was extended to construction firms with 20 or more employees. Starting from 2013, construction firms with 10-19 employees have also been included in the sample.

Availability period and frequency of the data

The databases are updated at the end of each run of the survey using the responses provided by the firms during the interviews.

The dataset containing the figures updated to the new reference year is made available to the users as soon as the survey report is published on the Bank of Italy's website (see the section 'Aggregated statistics and reference publications').

Data collection and regulatory references

The survey is conducted annually between February and May, with reference to economic activity at the end of the previous year. The interviews are led by the Bank of Italy's branches. Starting from the survey for 2010, all data are uploaded using a web-accessible application.

Aggregate statistics and reference publications

The main aggregate statistics are collected in tables available on the survey's webpage on the Bank of Italy's website.

The key findings of the survey are also published in a specific report of the 'Statistics Series' of the Bank of Italy and, alongside other information, in the *Annual Report*, the *Economic Bulletin* and the publications of the 'Regional Economies' series.

Databases structure and characteristics

The data collected through the Survey of Industrial and Service Firms and covering all the editions of the survey from 1984 onwards are stored in a single database (in csv format). Each edition of the survey is identified by the variable **annoril**, which indicates the year surveyed, i.e. the year prior to the one in which the data were actually collected (e.g. the data regarding the Survey for 2019, conducted in the early months of 2020, are identified by the variable annoril=2019).

Since the questionnaires administered to construction firms are very different from those administered to firms in industry excluding construction and in services, their respective databases

are also separate: the dataset containing only firms in industry excluding construction and in services is named 'indann_completo_csv', while the dataset containing only construction firms is named 'costr'.

In both datasets, each firm is assigned an identification code (the variable **ident**), which makes it possible to retrieve all the data available for firms that were included in the survey in multiple years. The code, generated randomly, is entirely unrelated to the variables that identify the firms and is used exclusively for longitudinal analyses. The **annoril/ident** pair identifies data on a given firm in a given year.

The databases contain some variables relating to the grossing-up of the sample estimates. Adopting a sampling weight allows for the alignment of the structure of the sample with that of the universe of firms according to the stratification variables; its use in the analyses is recommended to obtain unbiased estimates that can be grossed up to the reference population.

For the variables provided by the firms as ratios, or in any case without a scale factor, the weights applied should be such as to take account of the extent of the phenomenon. The databases also contain a number of variables relating to the classifications adopted for stratification on the basis of the sampling design. Please note that in the survey design, the geographical area is a post-stratification variable.

The financial data from the survey are expressed at current prices and in thousands of euros (except for the variables relating to wages, which are expressed in euros), also for the years prior to the introduction of the euro.

Since the survey for 2010, some focus variables, specifically indicated in the database, have been recorded only for half of the sample; the sections of the questionnaire that contain these variables are clearly marked with the letter 'A' or 'B', to indicate to which half of the sample they pertain.

The whole sample was divided into two subsets using a random mechanism, so as to maintain the representativeness of the two halves of the original sample in respect of the reference population. When analysing the variables observed on a subset, it is important to use grossing-up weights created specifically for that purpose (see the descriptions for pesoa and pesob here below) to make sure that the values are grossed up correctly in the subset examined.

Methodology and imputation of missing data

The survey follows a two-step stratified weighting procedure. In the first step, the combinations of business sector and size class are used as strata. Each firm is assigned an initial weight, given by the ratio of the number of firms in the stratum cell to the number of firms in the sample. The second step involves post-stratification, by using the raking technique, in order to take into account the geographical area in which the firm's administrative offices are located.

Where there are no answers for the main variables (e.g. planned investment expenditure, estimated turnover or employment), ratio estimators are used to impute data, setting the average number of the firm's employees in the reference year as denominator in order to capture the scale effect. In some cases, the firm's time series data are used for the reconstruction, in the form of individual effects. This method gives an estimate of a level per employee that is obtained by combining a general cross-section mean and an average calculated only based on the firm's time series data. The levels at times t and t+1 are reconstructed sequentially, by calculating average changes in appropriate cells of homogeneous firms. At that same moment, a new flag variable is created, making it possible to know if the value in the database was provided directly by the firm or if it was imputed: if the figure of the reference variable is derived from imputation, the flag takes the value 1, otherwise it is empty. Flag variables follow the naming convention f'X', where X is the name of the imputed variable (e.g. the imputation flag of the expected employment variable, v611m, is fv611m). However, the percentage of imputed data is very small for the key variables at the final and pre-final level. The forecast questions, particularly those relating to investment, tend to have a relatively higher non-response rate. The non-response rate for focus sections may also be high, depending on the difficulty of the questions.

Before the complete dataset is created, the responses undergo quality checks. The estimates published in the survey report, in the Statistics Series, refer to the databases as at the moment in which the calculations were performed and to the most recent data available regarding the population of firms. Small differences between the database used to prepare the estimates presented in the report and the one made available to prepare the calculations remotely can be the result of both subsequent revisions of the responses provided by the firms and of revisions in the weighting scheme due to a change in the reference population. The weighting scheme is revised

systematically, while any changes made to the responses provided in previous years are to be considered occasional.

For further details on the sampling design, on the construction of the weights and deflators and on all other methodological aspects, see the 'Methodological Notes' available on the Bank of Italy's website.

Variables description and characteristics

The description of the variables and the domain of the relevant attributes are contained in this excel file.

Variables description and characteristics (formats of the variables and codes/domains)

For confidentiality reasons, the databases are not searchable for questionnaire variables that would make it possible to identify the respondent firms. These variables are usually found in the first pages of the questionnaires and include: tax identification number, company name, subsidiary surveyed and group affiliation. Open text replies to questions of the type 'Other, specify' are also not searchable. Conversely, the databases contain some variables that are not included in the questionnaires (e.g. survey reference year, variable **annoril**) but which are useful for calculation purposes (see the full list in the Appendix).

How to cite the database

Banca d'Italia (2024): Survey of Industrial and Service Firms, July 2024 (1984-2023).

Appendix: Variables in the archives but not in the questionnaires

a) Classification variables with respect to the sector of economic activity[1]

Variable	*Values*	*Description*	*ATECO 2002*	*ATECO 2007*
settor11	*SS1*	Food industries, beverages and tobacco products	*DA*	*10, 11, 12*
	SS2	Textiles, clothing, and hide, leather and footwear products	*DB, DC*	*13, 14, 15*
	SS3	Coke manufacturing, chemical industry, rubber and plastics	*DF, DG, DH*	*19, 20, 21, 22*
	SS4	Processing of non-metallic minerals	*DI*	*23*
	SS5	Metal engineering industry	*DJ, DK, DL, DM*	*24, 25, 26, 27, 28, 29, 30, 33*
	SS6	Other manufacturing industries	*DD, DE, DN*	*16, 17, 18, 31, 32*
	SS7	Other industries excluding construction	*CA, CB, CE*	*05, 06, 07, 08, 09, 35, 36, 37, 38, 39*
	SS8	Wholesale and retail commerce	*G*	*45, 46, 47*
	SS9	Hotels and restaurants	*H*	*55, 56*
	SS10	Transport and communications	*I*	*49, 50, 51, 52, 53, 58, 59, 60, 61, 62, 63*
	SS11	Real estate activities, IT, etc.	*K*	*68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82*
indag3	*1*	Manufacturing industry	*D*	*C*
	2	Extractive Industries – Energy	*C, E*	*B, D, E*
	3	Services	*G, I, H, K*	*G, I, H, J, L, M, N*
indagine	*1*	Industry excluding construction	*C, D, E*	*C, B, D, E*
	2	Services	*G, I, H, K*	*G, I, H, J, L, M, N*

[1] Until 2009 the variables were obtained by aggregating two-letter groups (sub-sections) of the ISTAT ATECO 2002 classification; Starting from 2010, the classification variables by sector of economic activity are derived by aggregating two-digit groups (divisions) of the ISTAT ATECO 2007 classification. The variables in this table are contained only in the database 'indann_completo_csv.csv' and all the firms in the construction database have a single ATECO code (2002, 2007).

b) Classification variables with respect to the size class[2]

Variable	Values	Description
cldimet	0	20 - 49 employees
	1	50 - 99 employees
	2	100 - 199 employees
	3	200 - 499 employees
	4	500 - 999 employees
	5	≥ 1.000 employees
cc2	1	20 - 49 employees
	2	≥ 50 employees

c) Classification variables with respect to the geographical area [3]

Variable	Values	Description
areag4	1	North- West
	2	North-East
	3	Centre
	4	South and Islands
areag2	1	North, Centre
	2	South and Islands

d) Variables concerning the sample design and the weighting system

strato: Consisting of 66 combinations of settor11 and cldimet. It should be noted that firms with at least 5,000 workers have a weight of one and can be considered, individually for industry and services, as being part of two separate strata (i.e. 67 and 68).

poststrato: Consisting of 48 combinations of areag4, cc2 and a re-aggregation of the sectors of economic activity into 6 groups: 1) indag3=1; 2) indag3=2; 3) settor11=ss8; 4) settor11=ss9; 5) settor11=ss10; 6) settor11=ss11.

peso: Sample expansion weight: at the stratum and poststratum level, the sum of the weights is equal to the number of the reference population, separately for each year and does not take account of the panel dimension of the sample. A weight equal to 1 is assigned to firms with more than 5,000 employees ('self-representatives') and to a limited number of firms that are not considered representative of the stratum to which they belong.

pesoadd: Sample expansion weight: at the stratum and poststratum level, the sum of the weights is equal to the number of the reference population of employees, separately for each year and does not take account of the length of the sample (available from 2007). This weight is particularly relevant for the weighting of

[2] Up to the 2003 reference year, the size class refers to the number of workers at the end of the year; from 2004 onwards to the average number of workers during the year.

[3] For confidentiality reasons the classifications by region and province are not available, only those by macro-region.

categorical variables, as it allows for the different size scale of firms to be taken into account.

pesoa: Equivalent to peso for firms belonging to rotation 'A'. It should be used for the variables collected over half the sample (subsample A) (available from 2010).

pesoadda: Equivalent to pesoadd for firms belonging to rotation 'A'. It should be used for the variables collected over half the sample (subsample A) (available from 2010).

pesob: Equivalent to peso for firms belonging to rotation 'B'. It should be used for the variables collected over half the sample (subsample B) (available from 2010).

pesoaddb: Equivalent to pesoadd for firms belonging to rotation 'B'. It should be used for the variables collected over half the sample (subsample B) (available from 2010).

pesorisc: The pesorisc variable is obtained as the product of peso and a suitable scale factor in such a way that, year by year, it sums to the sample number.

popstr: Size of the population at the stratum level.

poppostr: Size of the population at the post-stratum level.

e) Classification variables of the share of turnover exported

Variable	Values	Description
a6	0	non-exporting firm
	1	less than 1/3 of turnover exported
	2	between 1/3 and 2/3 of turnover exported
	3	more than 2/3 of turnover exported
qexp	1	less than 1/3 of turnover exported or non-exporting firm
	2	between 1/3 and 2/3 of turnover exported
	3	more than 2/3 of turnover exported

f) Variables available at constant prices

The databases, limitedly to investment and turnover levels, contain variables expressed both at current and at constant prices. Constant prices refer both to the latest available reference year and to the reference year of each individual survey. They make it possible to calculate changes at constant prices for the year in which they were recorded, even after adding data from new editions of the survey. For firms in industry excluding construction and in services, the deflators are derived from those provided by the firms themselves, by aggregating the data at subsection and size class level.

Description	At current price	At constant prices referred to the most recent year	At constant prices referred to the survey reference year
Fixed investment t-1	v200	v200cos[(a)]	v200k[(a)]
Fixed investment t	v202	v202cos	v202k
Fixed investment t+1	v203	v203cos	v203k
Turnover t-1	v209	v209cos	v209k
Turnover t	v210	v210cos	v210k
Turnover t+1	v437	v437cos	v437k
Turnover exported t-1	v211	v211cos	v211k
Turnover exported t	v212	v212cos	v212k
Turnover exported t+1	v438	v438cos	v438k
Intangible investment[(b)] t-1	v810	v810cos	v810k
Intangible investment[(b)] t	v811	v811cos	v811k
Intangible investment[(b)] t+1	v812	v812cos	v812k

Note: (a) Available since 1985; (b) According to the ESA 2010, the item 'fixed assets' consists of expenditure for computer software, databases and mineral exploration and evaluation, but it excludes patents and trademarks.