

Description of the databases and examples from the Business Outlook Survey of Industrial and Service Firms

30 June 2023

For further information: statistiche@bancaditalia.it
www.bancaditalia.it/statistiche/index.html

1. General information

The data collected through the Business Outlook Survey of Industrial and Service Firms (hereinafter 'the Business Outlook Survey' – 'Sondtel'), conducted annually between September and October, cover all editions of the survey since **1993** for firms in industry excluding construction and in services and since **2007** for construction firms.

The sampling design of the Business Outlook Survey essentially mirrors the design of the Survey of Industrial and Service Firms ('Invind'), conducted in the spring of the same year. In principle, the firms contacted are the same for both surveys and any differences in the composition of the sample are due exclusively to sample attrition and to some firms possibly leaving the survey reference population in the summer.

For the Business Outlook Survey, the firms are asked mainly qualitative questions on the trends of the main economic variables and questions to revise the information recorded by the 'Invind' survey for the same year. Since a large portion of the questionnaire changes from year to year two different databases have been made available to the users (only for firms in industry excluding construction and in services):

- **db_sondstor**, which contains only the variables observed continuously over the years. The various years are identified by the variable **annoril**. Please note that in the Business Outlook Survey, the variable **annoril** indicates the year in which the Survey is conducted (e.g.: 2007 for the survey conducted between September and October 2007), while in the Survey of Industrial and Service Firms ('Invind') it indicates the reference year, i.e. the year prior to the one in which the data were actually collected (e.g. 2007 for the Survey conducted in early 2008);
- **db_sondxxxx**, which contains all the variables recorded in the year xxxx.

In the **db_sondxxxx** databases some of the variables used to record data on the same phenomenon may have been given different names and a partially different response scale over the years. In the **db_sondstor** dataset, variables with the same meaning have been attributed a unique code over time, and their scale of values has been harmonized, to ensure comparability over time. For example, the name of the variable 'Investments planned for the following year' is s3 in the db_sond1993 database and x3 in the db_sond1994 database, while in the db_sondstor database, the name of the variable is stg3 for all the years in which the variable was recorded. For construction firms, **db_sondcost** is the only database and it contains all the variables measured in the various years, with names that may change over time.

For a full list of all the variables recorded in the individual years and their recodification in the historical database, see the 'Variable description' database available on the [Bank of Italy's website](http://www.bancaditalia.it/statistiche/index.html).

Each firm is assigned an identification code (the variable **ident**), which makes it possible to retrieve information only from the **db_sondstor** database for firms that were included in the survey in multiple years. This code, generated randomly, is used exclusively for longitudinal analysis, cannot be tabulated and is entirely unrelated to the variables that identify the firms.

In the **db_sondstor** and **db_sondcost** databases, the **annoril/ident** pair identifies data on a given firm in a given year. If the same firm participates in the Survey of Industrial and Service Firms ('Invind') and in the Business Outlook Survey, it will have the same **ident** variable in all the databases.

Merely by way of example, the logical structure of the **db_sondstor** and **db_sondcost** databases (which cannot be visualized by the user) is as follows:

ANNORIL	IDENT	A5	STG1	STG3
1993	1	40	1	4
1993	2	190	2	9
...
2020	1	35	3	1
2020	2	240	9	2
...

At the same time, the logical structure of the **db_sond1993** and **db_sond2020** databases (which cannot be visualized by the user), is as follows:¹

DB_SOND1993

IDENT	a5m	s1	s3
1	40	1	4
2	190	2	9
...

DB_SOND2020

IDENT	a5m	p1	p3
1	35	3	1
2	240	9	2
...

Please note that the Business Outlook Survey is considered valid only if fully compiled. To this end, all the variables include among the response options '9 = do not know, do not wish to answer' and, where necessary, '8 = not applicable'. The missing values should thus not be interpreted as a firm refusing to provide a given response, but as questions asked only to a subset of firms; in some cases, for the group of firms that does not receive these questions, the relevant field can be populated with the figure '99'.

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;² its use in the analyses is recommended to obtain unbiased estimates that can be grossed up to the reference population.

2. Searchable variables in the databases

For confidentiality reasons, the databases are **not** searchable for questionnaire variables that would make it possible to identify the respondent firm. 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 but which are useful for calculation purposes. Besides the survey reference year (variable **annoril**), these include:

¹ The ident variables, which are shown here only to illustrate the relationship between the different databases, are not included in the annual datasets.

² Given that the size of the population becomes known with a one or two year lag, the weights are calculated using the size of the population in the strata of the most recent year. Small deviations between the actual number of units in the reference population and the sum of the weights are due to post-stratification.

a) Classification variables for sectors of economic activity³

Name	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

b) Classification variables for size classes⁴

Name	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 and over employees
cc2	1	20 - 49 employees
	2	50 and over employees

c) Classification variables for geographical areas⁵

Name	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 relating to the sampling design and weighting scheme⁶

³ Up to 2009 these variables were obtained by aggregating some of the subsections of the ATECO 2002 code; as of 2010, they are based on the first two digits of the ATECO 2007 code. The variables in this table are contained only in the databases of firms in industry excluding construction and in services, while all the firms in the construction database have a single ATECO code (2002, 2007).

⁴ The size classes are based on the end-of-year workforce until 2003 and on the average annual workforce thereafter.

⁵ For confidentiality reasons, only the classifications for the geographical macro-areas, and not those for the individual regions and provinces, are available.

⁶ 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' published on the Bank of Italy's website in conjunction with each of the [survey reports](#).

- 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.
- 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.
- 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.
- pesoadd:** Sample expansion weight: at the strato and poststrato 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).
- popstr:** Size of the population at the stratum level.
- poppostr:** Size of the population at the poststratum level.

a) *Classification variables for the share of turnover from exports*

Name	Values	Description
a6	1	non-exporting firm
	2	less than 1/3 of turnover exported
	3	between 1/3 and 2/3 of turnover exported
	4	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

3. Examples

3.1. Examples based on the software 'R'⁷

To obtain more rapid computing results, limit the number of variables in the datasets used to calculate the estimates. Please note that 'R' is a case-sensitive programming language.

All the examples assume that there is only one command per line and that the same command can extend over multiple lines if it is too long.

Example 1: frequency distribution

- This example shows how to tabulate, in the historical database and for all the years available, the frequency distribution of the response options for the variable **STG3** (investments planned for the following year), only for manufacturing firms with 50 or more employees (indag3==1).

##functions for data manipulation

```
library(dplyr)
library(data.table)
```

##functions for handling survey data

```
library(survey)
```

##functions for data summary

```
library(gtsummary)
```

##data Loading

```
dati <- fread("sondstor.csv")
```

⁷ 'R' is an opensource software environment for statistical data analysis; for more information on the programming language, visit the website <http://cran.r-project.org/>.

```

oggetto <- dati %>% filter(cc2>=2, indag3 == 1)

##create data frame containing variables of interest
oggetto <- oggetto %>% select(annoril, indag3, cc2, stg3, pesorisc)

out_svy <- svydesign(id= ~1, weights= ~pesorisc,
                  data=oggetto)
summary(out_svy)

##compute weighted relative frequencies
out_svy %>% tbl_svysummary(by = annoril,
                          percent = "column",
                          include = stg3,
                          statistic = list( stg3 ~ "{p}%" )) %>%
  modify_footnote(c(all_stat_cols()) ~ NA) %>%
  modify_header(c(all_stat_cols()) ~ "**{level}**", label = "")

```

Example 2: merging the two surveys

- This example shows how to merge the historical databases of the Business Outlook Survey and of the 'Invind' Survey to compare the investment plans observed during the 2021 edition of the Business Outlook Survey and the investments actually made that year (recorded as continuous data in the survey for 2021 and then discretized). We process the data only for the firms that participated in both surveys and which provided a valid response in the survey (other than option 9 = 'do not know, do not wish to answer').

```

##functions for data manipulation
library(dplyr)
library(data.table)

##data Loading
dati <- fread("sondstor.csv")

##create data frame containing variables of interest
##filter by year and type of firm
##by anno=2006 and stg3<>9
sond2021 <- dati %>% filter(annoril == 2021, stg3 !=9) %>%
  select(annoril, ident, stg3)

##data Loading (INVIND 2021)
dati <- fread("indann_completo_csv.csv")
##create data frame with data from the 2021 INVIND survey
invind2021 <- dati %>% filter(annoril == 2021) %>%
  select(annoril, ident, v200, v810, v202, v811, peso)%>%
  mutate(
    i0tot=v200+v810,
    i1tot=v202+v811,
    i0tot=if_else(i0tot==0, 0.1, i0tot),
    i1tot=if_else(i1tot==0, 0.1, i1tot),
    varinv=(i1tot/i0tot-1)*100,
    varinvd= cut(varinv,
                 breaks = c(-Inf, -10, -3, 3, 10,Inf),
                 labels = c(1,2,3,4,5)))

##create data frame with data from the 2021 INVIND survey
out_merge=inner_join(sond2021,invind2021) %>% select(stg3, peso, varinvd)

```

```

out_svy <- svydesign(id= ~1, weights= ~peso,
                   data=out_merge)
summary(out_svy)

##compute weighted relative frequencies
out_svy %>% tbl_svysummary(by = stg3,
                          percent = "cell",
                          include = varinvd,
                          statistic = list( stg3 ~ "{p}%", varinvd ~ "{p}%")) %>%
  modify_footnote(c(all_stat_cols()) ~ NA) %>%
  modify_header(c(all_stat_cols()) ~ "***{level}***", label = "stg3")

```

3.2. Examples based on the software ‘Stata’⁸

This software too is case-sensitive, so all commands must be lower case.

Example 1

- This example shows how to tabulate, in the historical database and for all the years available, the frequency distribution of the response options for the variable **STG3** (investments planned for the following year), only for manufacturing firms with 50 or more employees (indag3==1).

```

#delimit;

import delimited "sondstor.csv";

keep annoril indag3 cc2 stg3 pesorisc;

sort annoril;

by annoril: tabulate stg3 [iweight=pesorisc] if cc2>=2 & indag3==1;

```

Example 2

- This example shows how to merge the historical databases of the Business Outlook Survey and of the ‘Invind’ Survey to compare the investment plans observed during the 2021 edition of the Business Outlook Survey and the investments actually made that year (recorded as continuous data in the survey for 2021 and then discretized). We process the data only for the firms that participated in both surveys and which provided a valid response in the survey (other than option 9 = ‘do not know, do not wish to answer’).

```

#delimit;

/* Load data from Business Outlook survey 2021 */

import delimited "sond2021.csv";

keep annoril ident stg3 if annoril==2021;

sort ident;

save sond2021, replace;

clear;

/* Load data from Annual Survey 2021 */

import delimited "indann_completo_csv.csv";

keep annoril ident v200 v810 v202 v811 peso if annoril==2021;

```

⁸ ‘Stata’ is a registered trademark of StataCorp LP, 4905 Lakeway Drive, College Station, TX 77845 USA.

```

/* compute changes in investment */
generate i0tot=v200+v810;
generate i1tot=v202+v811;

/*substitute zeroes with small positive values to obtain a valid rate of change even if
the value of i0tot or both terms are zero */
replace i0tot=0.1 if i0tot==0;
replace i1tot=0.1 if i1tot==0;
generate varinv=(i1tot/i0tot-1)*100;

/* discretize continuous variable varinv */
generate varinvd=1 if varinv<-10;
replace varinvd=2 if (varinv>=-10 & varinv<-3);
replace varinvd=3 if (varinv>=-3 & varinv<=3);
replace varinvd=4 if (varinv>3 & varinv<=10);
replace varinvd=5 if varinv>10;

sort ident;
merge ident using sond2021;

/* keep only firms appearing in both datasets */
keep if _merge==3;
tabulate stg3 varinvd [iweight=peso] if stg3 !=9, cell;

```