

The Italy 2019 Enterprise Surveys Data Set

I. Introduction

This document provides additional information on the data collected in Italy between November 2018 and October 2019. The survey was part of a joint project of the European Bank for Reconstruction and Development (EBRD), the European Investment Bank (EIB) and the World Bank Group (WBG). The objective of the Enterprise Survey is to gain an understanding of what firms experience in the private sector.

As part of its strategic goal of building a climate for investment, job creation, and sustainable growth, the World Bank has promoted improving the business environment as a key strategy for development, which has led to a systematic effort in collecting enterprise data across countries. The Enterprise Surveys (ES) are an ongoing World Bank project in collecting both objective data based on firms' experiences and enterprises' perception of the environment in which they operate.

The ES currently cover over 180,000 firms in 150 countries, of which 142 have been surveyed following the standard methodology. This allows for better comparisons across countries and across time. Data are used to create statistically significant business environment indicators that are comparable across countries. The ES are also used to build a panel of enterprise data that will make it possible to track changes in the business environment over time and allow, for example, impact assessments of reforms.

This report outlines and describes the sampling design of the data, the data set structure as well as additional information that may be useful when using the data, such as information on non-response cases and the appropriate use of the weights.

II. Sampling Structure

The sample for 2019 Italy ES was selected using stratified random sampling, following the methodology explained in the *Sampling Note*¹. Stratified random sampling² was preferred over simple random sampling for several reasons³:

a. To obtain unbiased estimates for different subdivisions of the population with some known level of precision.

b. To obtain unbiased estimates for the whole population. The whole population, or universe of the study, is the non-agricultural economy. It comprises: all manufacturing sectors according to the group classification of ISIC Revision 3.1: (group D), construction sector (group F), services sector (groups G and H), and transport, storage, and communications sector (group I). Note that this definition excludes the following sectors: financial intermediation (group J), real estate and renting activities (group K, except sub-

¹ The complete text can be found at http://www.enterprisesurveys.org/~media/GIAWB/EnterpriseSurveys/Documents/Methodology/Sampling_Note.pdf

² A stratified random sample is one obtained by separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. (Richard L. Scheaffer; Mendenhall, W.; Lyman, R., "Elementary Survey Sampling", Fifth Edition).

³ Cochran, W., 1977, pp. 89; Lohr, Sharon, 1999, pp. 95

sector 72, IT, which was added to the population under study), and all public or utilities-sectors.

c. To make sure that the final total sample includes establishments from all different sectors and that it is not concentrated in one or two of industries/sizes/regions.

d. To exploit the benefits of stratified sampling where population estimates, in most cases, will be more precise than using a simple random sampling method (i.e., lower standard errors, other things being equal.)

e. Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This result is particularly true if measurements within strata are homogeneous.

f. The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.

Three levels of stratification were used in this country: industry, establishment size, and region. The original sample design with specific information of the industries and regions chosen is described in Appendix C.

Industry stratification was designed in the way that follows: the universe was stratified into four manufacturing industries and two services industries- Food and Beverages (ISIC Rev. 3.1 code 15), Fabricated Metal Products (ISIC code 28), Machinery & Equipment (ISIC code 29), Other Manufacturing (ISIC codes 16-27, 30-37), Retail (ISIC code 52) and Other Services (ISIC codes 45, 50, 51, 55, 60-64, and 72).

For the Italy ES, size stratification was defined as follows: small (5 to 19 employees), medium (20 to 99 employees), and large (100 or more employees).

Regional stratification for the Italy ES was done across the five NUTS1 regions: Northwest, Northeast, Center, South and Islands.

III. Sampling implementation

Given the stratified design, sample frames containing a complete and updated list of establishments as well as information on all stratification variables (number of employees, industry, and region) are required to draw the sample. Great efforts were made to obtain the best source for these listings.

Kantar Public Brussels, the main contractor, in collaboration with Lexis Ricerche implemented the Italy 2019 ES.

The sample frame consisted of a listing of establishments from Dun & Bradstreet. For establishments that were part of a multi-establishment firm and did not have establishment-level information on size, size information of the immediate headquarters was used to impute size bands as follows:

- If the immediate headquarters was small, the establishment was classified as small.
- If the immediate headquarters was medium or large, the establishment was classified as medium.

Table 1: Italy ES Sample Frame (Fresh)

		Food	Fabricated Metal Products	Machinery and Equipment	Other Manufacturing	Retail	Other Services	Grand Total
Northwest	Small (5-19)	60	180	100	360	240	660	3117
	Medium (20-99)	80	140	160	100	60	120	
	Large (100 or more)	150	220	267	80	80	60	
Northeast	Small (5-19)	60	120	80	320	160	580	2834
	Medium (20-99)	100	140	180	100	60	100	
	Large (100 or more)	141	173	280	80	100	60	
Center	Small (5-19)	80	80	80	340	260	480	2617
	Medium (20-99)	160	180	220	100	80	80	
	Large (100 or more)	36	60	40	160	121	60	
South	Small (5-19)	160	120	120	140	260	380	2657
	Medium (20-99)	220	240	280	100	100	60	
	Large (100 or more)	94	52	21	166	84	60	
Islands	Small (5-19)	280	240	211	120	280	180	2717
	Medium (20-99)	400	200	80	220	280	60	
	Large (100 or more)	14	16	6	27	43	60	
		2035	2161	2125	2413	2208	3000	13942

Source: Dun & Bradstreet

Necessary measures were taken to ensure the quality of the frame; however, the sample frame was not immune to the typical problems found in establishment surveys: positive rates of non-eligibility, repetition, non-existent units, etc.

The percentage of confirmed non-eligible units as a proportion of the total number of sampled establishments contacted for the survey was 4.5% (205 out of 4520 establishments)⁴.

Breaking down by industry and size, the following sample targets were achieved (based on the sampling information):

⁴ Based on out of target and ineligible contacts

Table 2: Achieved Interviews (Fresh)

		Food	Fabricated Metal Products	Machinery and Equipment	Other Manufacturing	Retail	Other Services	Grand Total
Northwest	Small (5-19)	3	9	5	18	12	33	162
	Medium (20-99)	4	7	8	5	3	6	
	Large (100 or more)	11	11	15	5	4	3	
Northeast	Small (5-19)	3	6	4	16	7	30	150
	Medium (20-99)	5	7	9	5	3	5	
	Large (100 or more)	12	12	14	4	5	3	
Center	Small (5-19)	4	4	4	17	13	24	150
	Medium (20-99)	8	9	13	5	4	4	
	Large (100 or more)	7	7	7	8	9	3	
South	Small (5-19)	8	6	6	7	13	17	148
	Medium (20-99)	11	12	14	5	5	3	
	Large (100 or more)	9	7	3	11	8	3	
Islands	Small (5-19)	14	12	14	6	14	9	150
	Medium (20-99)	20	10	0	11	14	3	
	Large (100 or more)	1	1	0	8	5	3	
	Medium and Large (20+)	0	0	5	0	0	0	
		120	120	121	131	119	149	760

IV. Data Base Structure:

The structure of the data base reflects the fact that 2 different versions of the survey instrument were used for all registered establishments. Questionnaires have common questions (*core* module) and respectfully additional manufacturing- and services-specific questions. The eligible manufacturing industries have been surveyed using the **Manufacturing** questionnaire (includes the *core* module, plus manufacturing specific questions). Retail firms have been interviewed using the **Services** questionnaire (includes the *core* module plus retail specific questions) and the residual eligible services have been covered using the **Services** questionnaire (includes the *core* module). Each variation of the questionnaire is identified by the index variable, *a0*.

All variables are named using, first, the letter of each section and, second, the number of the variable within the section, i.e. *a1* denotes section A, question 1 (some exceptions apply due to comparability reasons). Variable names preceded by the prefix ITA indicate questions specific to Italy while prefix “BM” or “BMG” indicate questions specific to Italy and other countries in Europe and Central Asia 2018/2019 and Middle East and North Africa 2019, therefore, they may not be found in the implementation of the rollout in other countries. All other suffixed variables are global and are present in all country surveys over the world. All variables are numeric with the exception of those variables with an “x” at the end of their names. The suffix “x” denotes that the variable is alpha-numeric.

There are 2 establishment identifiers, *idstd* and *id*. The first is a global unique identifier. The second is a country unique identifier. The variables *a2* (sampling region), *a6a* (sampling establishment’s size), and *a4a* (sampling sector) contain the establishment’s classification into the strata chosen for each country using information from the sample frame. The strata were defined according to the guidelines described above.

There are three levels of stratification: industry, size and region. Different combinations of these variables generate the strata cells for each industry/region/size combination. A distinction should be made between the variable *a4a* and *d1a2* (industry expressed as ISIC rev. 3.1 code). The former gives the establishment's classification into one of the chosen industry-strata based on the sample frame, whereas the latter gives the establishment's actual industry classification (four-digit code) based on the main activity at the time of the survey.

All of the following variables contain information from the sampling frame. They may not coincide with the reality of individual establishments as sample frames may contain inaccurate or outdated information. The variables containing the sample frame information are included in the data set for researchers who may want to further investigate statistical features of the survey and the effect of the survey design on their results.

-*a2* is the variable describing sampling regions

-*a6a*: coded using the same standard for small, medium, and large establishments as defined above.

-*a4a*: coded following the stratification by sector as defined above.

The surveys were implemented following a 2-stage procedure. Typically, first a screener questionnaire is applied over the phone to determine eligibility and to make appointments. Then a face-to-face interview takes place with the Manager/Owner/Director of each establishment. However, sometimes the phone numbers were unavailable in the sample frame, and thus the enumerators applied the screeners in person. The variables *a4b* and *a6c* contain the industry and size of the establishment from the screener questionnaire.

Note that there are variables for size (*l1*, *l6* and *l8*) that reflect more accurately the reality of each establishment. Advanced users are advised to use these variables for analytical purposes. Variables *l1* (number of permanent full-time workers at the end of the last complete fiscal year), *l6* (number of full-time seasonal workers employed during last complete fiscal year) and *l8* (average length of employment of full-time temporary employees during last complete fiscal year) were designed to obtain a more accurate measure of employment accounting for permanent and temporary employment. Special efforts were made to make sure that this information was not missing for most establishments.

The firms interviewed had several fiscal years. Most firms had January to December 2018 as their last complete fiscal year. Variables *a20m* (starting month of last complete fiscal year) and *a20y* (last complete fiscal year) can be used to obtain the last complete fiscal year for each firm.

For questions pertaining to monetary amounts, the unit is the Euro.

V. Universe

Universe figures for the number of establishments in each cell in Italy were obtained from Istat's 2015 ASIA registry (Registro statistico delle imprese attive).

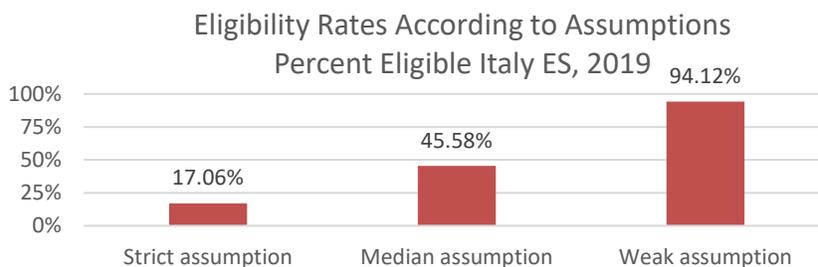
For Enterprise Surveys in which the sample is drawn from the same source as the universe, adjustments are made to the universe as a result of the impossibility to determine

eligibility for some of the establishments during the screening phase of the survey. In such cases, three sets of assumptions on establishment eligibility are used to adjust the universe and construct sample weights:

- Strict assumption: eligible establishments are only those for which it was possible to directly determine eligibility. The resulting weights are included in the variable *wstrict*.
- Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire, or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.
- Weak assumption: all establishments for which it was not possible to contact or that refused the screening questionnaire are assumed eligible. The resulting weights are included in the variable *wweak*.

For this survey, because the source of the sample was different than the source of universe figures, eligibility adjustments were not made, and *wstrict*, *wmedian* and *wweak* coincide.

The indicators computed for the ES website use the median weights. The following graph shows the different eligibility rates calculated for firms in the sample frame under each set of assumptions.



Weights for the probability of selection were computed using the number of completed interviews for each cell.

VI. Weights

Since the sampling design was stratified and employed differential sampling, individual observations should be properly weighted when making inferences about the population. Under stratified random sampling, unweighted estimates are biased unless sample sizes are proportional to the size of each stratum. With stratification the probability of selection of each unit is, in general, not the same. Consequently, individual observations must be weighted by the inverse of their probability of selection (probability weights or *pw* in Stata.)⁵

⁵ This is equivalent to the weighted average of the estimates for each stratum, with weights equal to the population shares of each stratum.

Due to non-response rates, some stratification cells were collapsed for the purposes of weighting, to preserve the representativeness of the sample. The following cells have been transformed: in the Islands region, Machinery & Equipment sector, medium and large firms were treated as one cell.

VII. Appropriate use of the weights

Under stratified random sampling, weights should be used when making inferences about the population. Any estimate or indicator that aims at describing some feature of the population should take into account that individual observations may not represent equal shares of the population.

However, there is some discussion as to the use of weights in regressions (see Deaton, 1997, pp.67; Lohr, 1999, chapter 11, Cochran, 1953, pp.150). There is not strong large-sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (stratum-specific coefficient): both simple OLS and weighted OLS are inconsistent under regular conditions. However, weighted OLS have the advantage of providing an estimate that is independent of the sample design. This latter point may be quite relevant for the ES as in most cases the objective is not only to obtain model-unbiased estimates but also design-unbiased estimates (see also Cochran, 1977, pp 200 who favors the used of weighted OLS for a common population coefficient.)⁶

From a more general approach, if the regressions are descriptive of the population then weights should be used. The estimated model can be thought of as the relationship that would be expected if the whole population were observed.⁷ If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, then, there is no reason to use weights.

VIII. Non-response

Survey non-response must be differentiated from item non-response. The former refers to refusals to participate in the survey altogether whereas the latter refers to the refusals to answer some specific questions. Enterprise Surveys suffer from both problems and different strategies were used to address these issues.

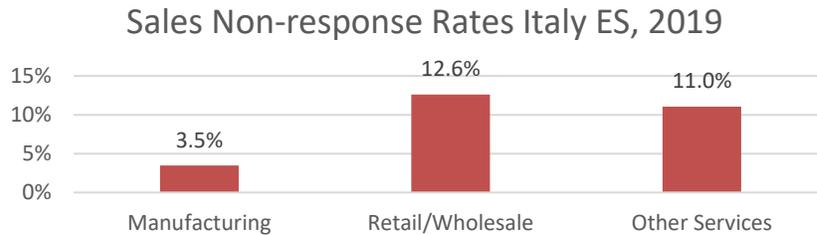
Item non-response was addressed by two strategies:

- a- For sensitive questions that may generate negative reactions from the respondent, such as corruption or tax evasion, enumerators were instructed to collect the refusal to respond (-8) as a different option from don't know (-9).
- b- Establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of

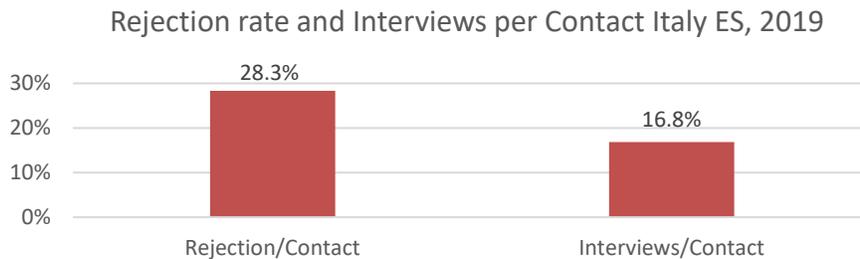
⁶ Note that weighted OLS in Stata using the command regress with the option of weights will estimate wrong standard errors. Using the Stata survey specific commands svy will provide appropriate standard errors.

⁷ The use weights in most model-assisted estimations using survey data is strongly recommended by the statisticians specialized on survey methodology of the JPSM of the University of Michigan and the University of Maryland.

low response. The following graph shows non-response rates for the sales variable, d2, by sector. Please, note that for this specific question, refusals were not separately identified from “Don’t know” responses.



As the following graph shows, the number of interviews per contacted establishments was 0.17.⁸ This number is the result of two factors: explicit refusals to participate in the survey, as reflected by the rate of rejection (which includes rejections of the screener and the main survey) and the quality of the sample frame, as represented by the presence of ineligible units. The share of rejections per contact was 0.28.



Details on the rejection rate, eligibility rate, and item non-response are available at the level strata. This report summarizes these numbers to alert researchers of these issues when using the data and when making inferences. Item non-response, selection bias, and faulty sampling frames are not unique to Italy. All enterprise surveys suffer from these shortcomings, but in very few cases they have been made explicit.

References:

Cochran, William G., *Sampling Techniques*, New York, New York: John Wiley & Sons, 1977.

Deaton, Angus, *The Analysis of Household Surveys*, Baltimore, Maryland: Johns Hopkins University Press, 1998.

Levy, Paul S. and Stanley Lemeshow, *Sampling of Populations: Methods and Applications*, New York, New York: John Wiley & Sons, 1999.

Lohr, Sharon L. *Sampling: Design and Techniques*, Boston, Massachusetts: Brookes/Cole, 1999.

Scheaffer, Richard L.; Mendenhall, W.; Lyman, R., *Elementary Survey Sampling*, Fifth Edition, 1996.

⁸ The estimate is based on the total no. of firms contacted including ineligible establishments.

Appendix A

Status Codes Enterprise Survey (ES) :

61	Screening in process	14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	61
771	Eligible	1. Eligible establishment (Correct name and address)	766
		2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	3
		3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	0
		4. Eligible establishment (Moved and traced)	2
		16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.)	0
1270	Screener refusal	13. Refuses to answer the screener	1270
169	Ineligible	5. The establishment has less than 5 permanent full time employees	13
		616. The firm discontinued businesses - (Establishment went bankrupt)	0
		618. The firm discontinued businesses - (Original establishment disappeared and is now a different firm)	0
		619. The firm discontinued businesses - (Establishment was bought out by another firm)	0
		620. The firm discontinued businesses - (It was impossible to determine for what reason)	4
		621. The firm discontinued businesses - (Other)	3
		71. Ineligible legal status: not a business, but private household	0
		72. Ineligible legal status: cooperatives, non-profit organizations, etc.	148
36	Out of Target	8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	1
		151. Out of target - outside the covered regions	0
		152. Out of target - moved abroad	0
		153. Out of target - Not registered with Statistical Authority	0
		154. Out of target - establishment is HQ without production or sales of goods or services	4
		155. Out of target - establishment was not in operation for the entirety of last fiscal year	1
		156. Duplicated firm within the sample	8
2213	Unobtainable	157. Out of target - location that is not HQ and does not have financial statements prepared separately	23
		91. No reply after having called in different days of the week and in different business hours	854
		92. Line out of order	30
		93. No tone	224
		94. Phone number does not exist	1083
		10. Answering machine	4
		11. Fax line- data line	15
		12. Wrong address/ moved away and could not get the new references	3
4520	Total contacted		

Response Outcomes : Italy ES 2019 :

Target and totals	Sample target	760
	Sample target completion rate	100.0%
	Total contacts available in frame	13942
	Total contacts issued	5135
	Total contacts contacted	4520

Screening phase	Screening in process	61
	Eligibles	771
	Screener refusal	1270
	Ineligible + out of target	205
	Unobtainable	2213
Interview phase (only if eligible)	Complete interviews without extra module	0
	Complete interviews with extra module	760
	Eligible in process + incomplete interviews	0
	Interview refusal	11

Percent breakdown (relative to total contacted)	Screening in process rate	1.3%
	Screener refusal rate	28.1%
	Ineligible + out of target rate	4.5%
	Unobtainable rate	49.0%
	Interview conversion rate	16.8%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	0.2%

Appendix B

Universe

		Food	Fabricated Metal Products	Machinery and Equipment	Other Manufacturing	Retail	Other Services	Grand Total
Northwest	Small (5-19)	2767	6986	3780	15783	11668	46900	111381
	Medium (20-99)	732	1854	1712	5491	2019	8484	
	Large (100 or more)	161	163	336	1149	329	1067	
Northeast	Small (5-19)	2785	5093	3312	14089	9528	41105	95615
	Medium (20-99)	839	1581	1581	4966	1693	6712	
	Large (100 or more)	174	141	373	861	150	632	
Center	Small (5-19)	2256	2249	899	12763	9782	34137	74797
	Medium (20-99)	345	534	369	3134	1512	5356	
	Large (100 or more)	46	44	57	483	151	680	
South	Small (5-19)	2623	1466	509	6467	9017	27185	55932
	Medium (20-99)	545	375	174	1699	845	4211	
	Large (100 or more)	61	44	18	252	55	386	
Islands	Small (5-19)	1270	417	97	1629	4664	11795	22900
	Medium (20-99)	178	69	29	332	496	1653	
	Large (100 or more)	9	6	7	50	34	165	
		14791	21022	13253	69148	51943	190468	360625

Note: The sampling frame used and the universe are from separate sources.

Appendix C: Original Sample Design

Original Sample Design (Fresh)

		Food	Fabricated Metal Products	Machinery and Equipment	Other Manufacturing	Retail	Other Services	Grand Total
Northwest	Small (5-19)	3	9	5	18	12	33	160
	Medium (20-99)	4	7	8	5	3	6	
	Large (100 or more)	11	11	14	4	4	3	
Northeast	Small (5-19)	3	6	4	16	8	29	150
	Medium (20-99)	5	7	9	5	3	5	
	Large (100 or more)	12	12	14	4	5	3	
Center	Small (5-19)	4	4	4	17	13	24	150
	Medium (20-99)	8	9	11	5	4	4	
	Large (100 or more)	7	7	9	8	9	3	
South	Small (5-19)	8	6	6	7	13	19	150
	Medium (20-99)	11	12	14	5	5	3	
	Large (100 or more)	9	7	3	11	8	3	
Islands	Small (5-19)	14	12	14	6	14	9	150
	Medium (20-99)	20	10	4	11	14	3	
	Large (100 or more)	1	1	1	8	5	3	
		120	120	120	130	120	150	760