Enterprise Survey, Micro, India 2022

I. Introduction

This document provides information on the Enterprise Surveys of Micro firms (ESM) conducted by the World Bank Group's (WBG) Enterprise Analysis Unit (DECEA) in India. The survey covers nine cities: Hyderabad, Telangana; Jaipur, Rajasthan; Kochi, Kerala; Ludhiana, Punjab; Mumbai, Maharashtra; Sehore, Madhya Pradesh; Surat, Gujarat; Tezpur, Assam; and Varanasi, Uttar Pradesh. The fieldwork was implemented by Nielsen (India) Pvt Ltd, and the data was collected between December 2021 and March 2022. The ESM in cities in India was made possible thanks to the cooperation with Omidyar Network India.

The primary objectives of the ESM are to: i) understand demographics of the micro enterprises in the covered cities, ii) describe the environment within which these enterprises operate, and iii) enable data analysis based on the samples that are representative at each city level.

This report briefly describes the sampling design of the data, the structure of the dataset as well as additional information that may be useful for data users.

II. Universe and Sampling Frame

The universe of ESM includes formally registered businesses in the sectors covered by the ES and with less than five employees. The definition of formal registration can vary by country. The universe table for each of the nine cities covered by ESM in India was obtained from the 6th Economic Census (EC) of India (conducted between January 2013 and April 2014), which has its own well-defined definition of registration. Generally, this entails registration with any central/government agency, under Shops & Establishment Act, Factories Act etc.

In terms of sectors, the survey covers all non-agricultural and non-extractive sectors. In particular, according to the group classification of ISIC Revision 4.0, it includes: all manufacturing sectors (group D), construction (group F), wholesale and retail trade (group G), transportation and storage (group H), accommodation and food service activities (group I), a subset of information and communications (group J), some administrative and support service activities (codes 79) and other service activities (codes 95). Notably, the ESM universe excludes the following sectors: financial and insurance activities (group K), real estate activities (group L), and all public or utilities-sectors.

Due to difficulties obtaining reliable contact information from the full universe of establishments included in the ESM, the sampling frame was constructed through the Enterprise Survey of the Informal Sector (ESIS) conducted in the same nine cities in India. For detailed information on the methodology of that survey, please consult the corresponding implementation report. To construct the sampling frame for ESM, only the starting blocks of ESIS were used.[1] Notably, the definition of formal registration used in the construction of the ESM sampling frame (and in ESIS in general) was different from the one used in the EC (source of the universe table). In particular, an enterprise is considered formally registered if it either has a business PAN or a GST number. Table 1 report sizes of the ESM universe and frame for the India 2022 ESM.

[1] The contact information of micro enterprises was not collected from the expansion blocks of the ESIS because of the difficulty of estimating the corresponding ESM sampling weights.

Table 1: India ESM 2022 Universe and Sample Frame

	Universe			Sample Frame		
	Manufacturing	Services	Total	Manufacturing	Services	Total
Hyderabad	54,567	202,612	257,179	65	521	586
Jaipur	22,230	57,457	79,687	52	215	267
Kochi	8,967	18,141	27,108	46	724	770
Ludhiana	25,153	66,244	91,397	64	464	528
Mumbai	79,928	214,574	294,502	165	1574	1739
Sehore	1,482	4,386	5,868	85	437	522
Surat	62,565	126,497	189,062	79	639	718
Tezpur	1,994	12,379	14,373	87	442	529
Varanasi	38,392	58,850	97,242	10	96	106
TOTAL	295,278	761,140	1,056,418	653	5112	5765

Source: 6th Economic Census, and India 2022 ESIS, starting blocks.

III. Sampling Structure

The sample for Enterprise Survey of Micro firms in India 2022 was selected using stratified random sampling, following the methodology explained in the *Sampling Note*.[2] Stratified random sampling was preferred over simple random sampling for several reasons, including:[3]

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

b. 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.

c. 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.)

d. 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.

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

Two levels of stratification were used in this survey: industry and region. For stratification by industry, two groups were used: Manufacturing (combining all the relevant activities in ISIC Rev. 4.0 codes 10-33) and Services (remainder of the universe, as outlined above). Regional stratification was done across nine cities included in the study, namely: Hyderabad, Jaipur, Kochi, Ludhiana, Mumbai, Sehore, Surat, Tezpur and Varanasi.

[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). ES Sampling Note containing more details is available at: https://www.enterprisesurveys.org/content/dam/enterprisesurveys/documents/methodology/Sampling_Note-Consolidated-2-16-22.pdf

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

IV. Survey Implementation

The fieldwork was implemented by Nielsen (India) Pvt Ltd, and the data were collected between December 2021 and March 2022. The script was written in Survey Solutions. While the sampling frame was constructed shortly before the ESM implementation, it was not immune to the typical problems found in establishment surveys, such as positive rates of non-eligibility. The percentage of confirmed non-eligible units as a proportion of the total number of contacted establishments for the survey was 3% (36 out of 1203 establishments). Table 2 shows the number of interviews achieved in each stratification region and sector. Summary of status codes of all the establishments contacted for the survey is in the Appendix.

Table 2: Achieved Interviews

	Manufacturing	Services	Total
Hyderabad	51	57	108
Jaipur	46	61	107
Kochi	34	66	100
Ludhiana	60	50	110
Mumbai	54	59	113
Sehore	74	48	122
Surat	49	64	113
Tezpur	59	64	123
Varanasi	8	94	102
	435	563	998

V. Database Structure

The variable naming reflects whether or not the respective variable is also present in the ES or ESIS databases, for cross-size/registration-status comparability. In particular, all variables the first letter of which matches the section of the questionnaire are also present in the standard ES (not only in India but including most other ES). The variables starting with letter *i* are also present in the ESIS. All the variables that are specific to ESM start with the letter *m*. The variables with *msc* in the name are specific to the India ESM, and thus may not be found in the implementation of the rollout in other countries. All variables are numeric with the exception of those variables with an *x* at the end of their names, which are 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. As these stratification variables contain information from the sampling frame, they may not coincide with the reality of individual establishments as more accurate information is collected through the detailed ESM interview.

As noted above, there are two levels of stratification: industry (variable *a4a*) and region (*a2*). Different combinations of these variables generate the strata cells for each industry/region combination (variable *strata*). A distinction should be made between the variable a4a and d1a2_v4/d1a2 (industry expressed as ISIC rev. 4.1 and 3.1, respectively). 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.

The survey was implemented following a two-stage procedure. Sometimes, 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_v4* and *a6c* contain the industry and size of the establishment from the screener questionnaire. Note that there are variables for size (section L) that reflect more accurately the reality of each establishment. Users are advised to use these variables for analytical purposes.

Most of the questions in the ESM instrument refer to the information about the last completed fiscal year. For India 2022 ESM, this date was March 31, 2021. For questions pertaining to monetary amounts, the unit is the Indian rupee.

VI. Sampling 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).[4]

As noted above, the universe table and the sampling frame used in the India 2022 ESM are from different sources. Consequently, the eligibility adjustments that is standard for all ES survey was not applied and the base weights were used. This means that the sampling weight for each stratum was calculated as the ratio between the universe (given in Table 1) and the number of achieved interviews (given in Table 2).

VII. Appropriate Use of the Sampling 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

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

estimates but also design-unbiased estimates (see also Cochran, 1977, pp 200 who favors the used of weighted OLS for a common population coefficient).[5]

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.[6] 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

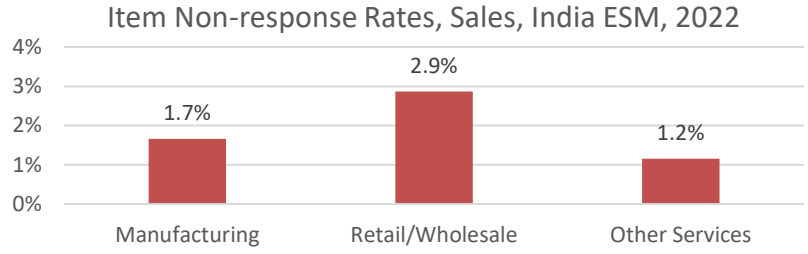
Survey non-response must be differentiated from item non-response. The former refers to refusals to participate in the survey altogether whereas the latter refers to the refusals to answer some specific questions. ESM suffer from both problems and different strategies were used to address these issues.

The 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 low response.

The graph below shows item 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.

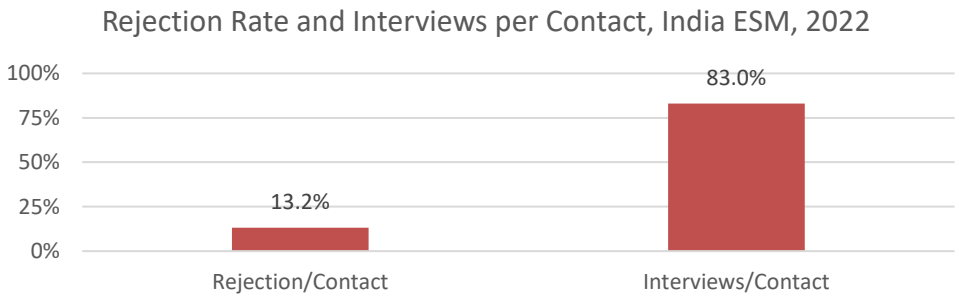


The survey non-response is show on the graph below, with the number of interviews per contacted establishments at 0.83.[7] This number is the result of two factors: explicit refusals to participate in the survey, as reflected by the rate of rejection (which includes rejections of the screener and the main survey) and the quality of the sample frame, as represented by the presence of ineligible units. The share of rejections per contact was 0.13.

[5] 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.

[6] 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.

[7] The estimate is based on the total number of firms contacted including ineligible establishments.



Details on the rejection rate, eligibility rate, and item non-response are available upon request at the level strata. The Appendix includes further details to alert researchers about these issues when using the data and when making inferences. Item non-response, potential selection bias, and faulty sampling frames are not unique to India as all surveys suffer from these shortcomings.

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. Samping: Design and Techniques, Boston, Massachusetts: Brookes/Cole, 1999.

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

Appendix

Status Codes:

0	**Screening in process**	14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0
1133	**Eligible**	1. Eligible establishment (Correct name and address)	1076
		2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	48
		3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	9
		4. Eligible establishment (Moved and traced)	0
		16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.)	0
25	**Screener refusal**	13. Refuses to answer the screener	25
36	**Ineligible**	5. The establishment has less than 5 permanent full time employees	4
		616. The firm discontinued businesses - (Establishment went bankrupt)	2
		618. The firm discontinued businesses - (Original establishment disappeared and is now a different firm)	2
		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)	0
		621. The firm discontinued businesses - (Other)	8
		71. Ineligible legal status: not a business, but private household	0
		72. Ineligible legal status: cooperatives, non-profit organizations, etc.	4
		8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	16
4	**Out of Target**	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	4
		154. Out of target - establishment is HQ without production or sales of goods or services	0
		155. Out of target - establishment was not in operation for the entirety of last fiscal year	0
		156. Duplicated firm within the sample	0
5	**Unobtainable**	91. No reply after having called in different days of the week and in different business hours	0
		92. Line out of order	0
		93. No tone	0
		94. Phone number does not exist	0
		10. Answering machine	0
		11. Fax line- data line	0
		12. Wrong address/ moved away and could not get the new references	5
1203	**Total contacted**		

Response Outcomes:

Target and totals	Sample target	1000
	Sample target completion rate	99.8%
	Total contacts available in frame	5765
	Total contacts issued	1408
	Total contacts contacted	1203

Screening phase	Screening in process	0
	Eligibles	1133
	Screener refusal	25
	Ineligible + out of target	40
	Unobtainable	5
Interview phase (only if eligible)	Complete interviews without extra module	998
	Complete interviews with extra module	0
	Eligible in process + incomplete interviews	0
	Interview refusal	134

Percent breakdown (relative to total contacted)	Screening in process rate	0.0%
	Screener refusal rate	2.1%
	Ineligible + out of target rate	3.3%
	Unobtainable rate	0.4%
	Interview conversion rate	83.0%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	11.1%