Republic of Albania Enterprise Surveys Data Set

1. Introduction

This document provides additional information on the data collected in Albania from 13 December 2007 to 24 March 2008 as part of the Enterprise Survey, an initiative of the World Bank.

The objective of the Enterprise Surveys is to obtain feedback from enterprises in client countries on the state of the private sector as well as 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.

Through interviews with firms in the manufacturing and services sectors, the Enterprise Survey data provides information on the constraints to private sector growth and is used to create statistically significant business environment indicators that are comparable across countries.

The report describes the sampling design of the survey, the structure of the dataset and additional information that may be useful when using the data, including information on non-response rates, the calculation of sample weights and country-specific factors that may have affected survey implementation.

2. Sampling Structure

The whole population, or the universe, covered in the Enterprise Surveys is the non-agricultural economy. It comprises: all manufacturing sectors according to the ISIC Revision 3.1 group classification (group D), construction sector (group F), services sector (groups G and H), and transport, storage, and communications sector (group I). Note that this population definition excludes the following sectors: financial intermediation (group J), real estate and renting activities (group K, except sub-sector 72, IT, which was added to the population under study), and all public or utilities-sectors.

The sample for the Republic of Albania was selected using stratified random sampling, following the methodology explained in the Sampling Manual. Stratified random sampling[1] was preferred over simple random sampling for several reasons[2]:

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

[1] 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).

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

sectors: financial intermediation (group J), real estate and renting activities (group K), 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: firm sector, firm size, and geographic region. The original sample design, with specific targets for these strata, is included in the attached Excel file (Sampling Report.xls.)

Industry stratification was designed in the way that follows: the universe was stratified into 1 manufacturing sector (including several industries), 1 services industry -retail-, and one residual sector as defined in the sampling manual. 304 interviews were completed in total, out of an original target of 360 interviews. The main constraint to reach the target interviews was the Universe size and composition, which proved to be smaller than originally expected. Particularly, firms with more than five employees in the Services sector were scarce. Firms from sector 51 (Wholesale) were issued to compensate the shortfall in the Services sector 52 (Retail). The majority of the relevant information, including the accounting data was obtained and entered into the data base. The Productivity section had a high non-response rate on average, reaching between 20 - 25%, depending on the questionnaire. Even if call backs were done to complete the section, the response rate could not be improved by much.

Size stratification was defined following the standardized definition used for the Enterprise Surveys: micro (1 to 4 employees), small (5 to 19 employees), medium (20 to 99 employees), and large (more than 99 employees). For stratification purposes, the number of employees was defined on the basis of reported permanent full-time workers.

Regional stratification was defined in terms of the geographic regions with the largest commercial presence in the country: Tirana, Durres, Elbasan, Fier, Vlora were the four metropolitan areas selected in Albania.

3. Sampling implementation

It was not possible to obtain a usable frame for Albania. Therefore, the design returned to first principles, using a blocks enumeration methodology. Detailed maps of major cities were obtained. These were from aerial mappings projected to a usable scale. They served as the basis of a multi-stage approach. Firstly each city (region) was divided into ‘blocks’ and then the blocks were classified into strata defined by the predominant spatial use, using local knowledge. The classifications used for the blocks included industrial, commercial, commercial/residential (mixed), and residential coding. The accuracy of the classification was tested using 30 pilot blocks. That test proved successful. Subsequently another 328 blocks were selected and enumerated; building by building, floor by floor. Each separate unit was identified, classified as to use and in the case of business establishments further details collected as to employee numbers, activity, name, and phone number. This enumeration of a total of 358 blocks was then employed to project to universe totals by reference to the screening results and the number of blocks in each stratum. The establishments enumerated in those blocks were then used as the frame for the selection of a sample with the aim of obtaining interviews at 360 establishments with five or more employees. In addition the World Bank requested interviews at 120 small manufacturing establishments with less than five employees, to be delivered separately as an additional survey. That target was subsequently reduced to 80 as only some 180 small manufacturing establishments had been enumerated. Disproportionate methods were used to reduce the variance of estimates.

The quality of the frame was assessed at the onset of the project. The frame proved to be useful though it showed positive rates of non-eligibility, repetition, non-existent units, etc. These problems are typical of establishment surveys, but given the impact these inaccuracies may have on the results, adjustments were needed when computing the appropriate weights for individual observations. The percentage of confirmed non-eligible units as a proportion of the total number of contacts to complete the survey was 6.8% (29 out of 425 establishments).

Sample selection was carried out by the TNS team in London using the data obtained from the block enumeration. The selections for Albania were augmented by additional selections from enterprises interviewed during the BEEPS survey in 2005 and a 'Large Taxpayers' database obtained by the local agency. To reduce non-response bias the samples was drawn in matched replicates so that each sampled establishment had at least one matched substitute (where sample available) in the event of non-contact or refusal.

Local Agency team involved in the study:

Local Agency	Name: IDRA Research & Consulting Country: Albania Membership of international organisation: ESOMAR Activities since: 2001
Name of Project Manager	Auron Pasha
Name and position of other key persons of the project:	Florian Babameto – Coordinator Adela Gjergjani – Fieldwork coordinator Rozeta Koci – Fieldwork coordinator Enton Coka – Data entry and quality control
Enumerators involved:	Enumerators: 50 enumerators in charge of the blocks enumeration and 50 interviewers in the second phase. Recruiters: the interviewers were also in charge

	of the recruitment
Other staff involved:	Fieldwork Coordinators: 2 people Editing: 1 supervisor Data Entry: 4 people

Sample Frame:

Characteristic of sample frame used:	
Source:	Block Enumeration Sample Frame + Albania's Large tax payer's data Base + BEEPS 2005 panel.
Year:	2005, 2007 - 2008
Comments on the quality of sample frame:	The retail sector in Albania is mainly composed by small companies. This was first noticed when analysing the results from the blocks enumeration and confirmed later during fieldwork. In addition, Manufacturing firms on the ground proved to be fewer than originally estimated prior to the beginning of the survey. From the BEEPS 2005 panel sample, many companies had changed size in the past years, and they and this was only confirmed after the screener was done.

Sample Frame Albania

	Classification			
Employees	Manufacturing	Retail	Other	Grand Total
5 to 19	84	140	66	290
20 to 99	55	21	22	98
100 and over	22	1	8	31
NA	17	104	21	142
Grand Total	178	266	117	561

Source: Block Enumeration conducted by the local Agency
Year: 2007

Sectors included in the Sample:

Original Sectors	Manufacture, Services, Residual
Added Sectors	Sector 51 was used as a top up for the Services sector

Sample:

Comments/ problems on sectors and	Since the panel from the BEEPS 2005 and the Large Taxpayers sample frame were used as a top up for the blocks

regions selected in the sample:	enumeration sample frame, the 2008 Enterprise Survey was not only conducted in the original sample design regions as defined by the World Bank, but it was also spread in other major towns such as Korca, Gjirokastra and Shkodra.
Comments on the response rate:	Response rate we achieved was above 50%, resulting from major efforts done to convince businesses to participate in the survey.
Comments on the sample design:	
Other comments:	None

Fieldwork:

Date of Fieldwork	13/12/2007 – 24/03/2008
Country	Albania
Interview number with more than five employees	Manufacturing: 110 Services: 81 Core: 113
Problems found during fieldwork:	The major problem during the field work was fixing the appointments with the firms. In order to get the interviews, firms had to be contacted several times.
Other observations:	None

4. Data Base Structure:

The structure of the data base reflects the fact that 3 different versions of the questionnaire were used. The basic questionnaire, the Core Module, includes all common questions asked to all establishments from all sectors (manufacturing and services). The second expanded variation, the Manufacturing Questionnaire, is built upon the Core Module and adds some specific questions relevant to the sector. The third expanded variation, the Services Questionnaire, is also built upon the Core Module and adds to the core specific questions relevant to retail. 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*. Variable names preceded by a prefix "*al"* are specific to the Republic of Albania and, therefore, they may not be found in the implementation of the Enterprise Survey 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 3 establishment identifiers, *idstd*, *idu*, and *id*. The first is a global unique identifier. The second is a regional unique identifier, and *the* third one 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.

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. These variables generate the strata cells for each industry/region/size combination. 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 the sampling regions
> -*a6a*: coded using the definition for micro, small, medium, and large establishments as discussed above. The code *-9* was used to indicate units for which size was undetermined in the sample frame.
> -*a4a*: coded using ISIC codes for the industries that comprise the manufacturing, services, and residual categories used in the stratification. These codes include most manufacturing industries (15 to 37), and retail, and IT for services (52, and 72 respectively). All establishments within the 'other manufacturing' stratum were coded with a4a=2.

Note that these variables may not coincide with reality for some establishments as sample frames may contain information that is later found to be inaccurate.

The surveys were implemented following a two stage procedure. In the first stage a screener questionnaire was administered over the phone to determine sampled establishment's eligibility for the survey and to make appointments; in the second stage, a face-to-face interview took place with the Manager/Owner/Director of each establishment. The variables *a4b* and *a6b* contain the industry and size of the establishment from the screener questionnaire. Variables *a8* to *a11* contain additional information that was collected in the screening phase.

The main questionnaire contains variables for location (*a3x*), industry (*d1a2*), and number of employees (*l1*, *l6* and *l8*) that more accurately reflect describes the characteristics of establishments than the information provided on these variables in the sample frame or the screener.

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 industry-strata based on information available from the sample frame, whereas variable *d1a2* indicates the actual ISIC code of the main output of the establishment as answered by the interviewee. This variable is probably the most accurate variable with which to classify establishments by activity.

Variable *a3x* indicates the actual location of the establishment. There may be divergences between the location in the sampling frame and the actual location, as establishments may be listed in one place on the sample frame but the actual physical location is in another place.

Variables *l1*, *l6* and *l8* provide a more accurate measure of employment and account for both permanent and temporary employment. Special efforts were made to make sure that this information was not missing for most establishments.

5. Universe Estimates

For Albania, ratios of the total numbers of blocks of each type to the totals enumerated were formed. Those ratios were then applied to the eligible establishments enumerated to provide universe estimates. Appendix C shows the overall estimates of the numbers of establishments based on the block ratios.

6. Weights

Since the sampling design was stratified and employed differential sampling of the strata, 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 *pa* in Stata.)[3]

As the sample frame was built based on a blocks enumeration methodology, only one set of weights for each cell was computed using the strict on establishment eligibility. Under the strict assumption eligible establishments are only those for which it was possible to directly determine eligibility.

A pair of weight sets was calculated. The first set of estimates calculated proportions using the raw sample count for each cell. However, the achieved sample numbers in many cells were small. Hence, those eligibility rates, and the adjusted universe cells projections, are subject to relatively large sampling variations. Therefore a second set of more robust estimates (collapsed weights) was also produced. These estimates made use of the multiples of the relative eligibility rates for each industry, size, and region. Those relative rates were based on much larger samples than the individual cells and thus produced values with smaller sampling variations. The data sets include only these robust weights.

7. Appropriate use of weights

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

As discussed above, 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 on the proper 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 has the advantage of providing an estimate that is independent of the sample design. This latter point may be quite relevant for the Enterprise Surveys 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).

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.[4] If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, there is no reason to use weights.

8. Non-response

The Enterprise Surveys, along with all other surveys, suffer from both survey non-response and 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. Different strategies were used to address these issues.

Survey non-response was addressed by maximizing efforts to contact establishments that were initially sampled. When the survey frame was extracted from the sampling frame, several establishments with the same strata characteristics were randomly selected for each interview and each establishment was assigned a preference number.[5] Substitutions of replacement establishments were made in order to help achieve targets on the number of interviews for each stratum. Extensive efforts were made to complete interviews with each first preference establishment before contact with a replacement establishment was allowed. At least four attempts were made to contact each sampled establishment for an

[4] 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.

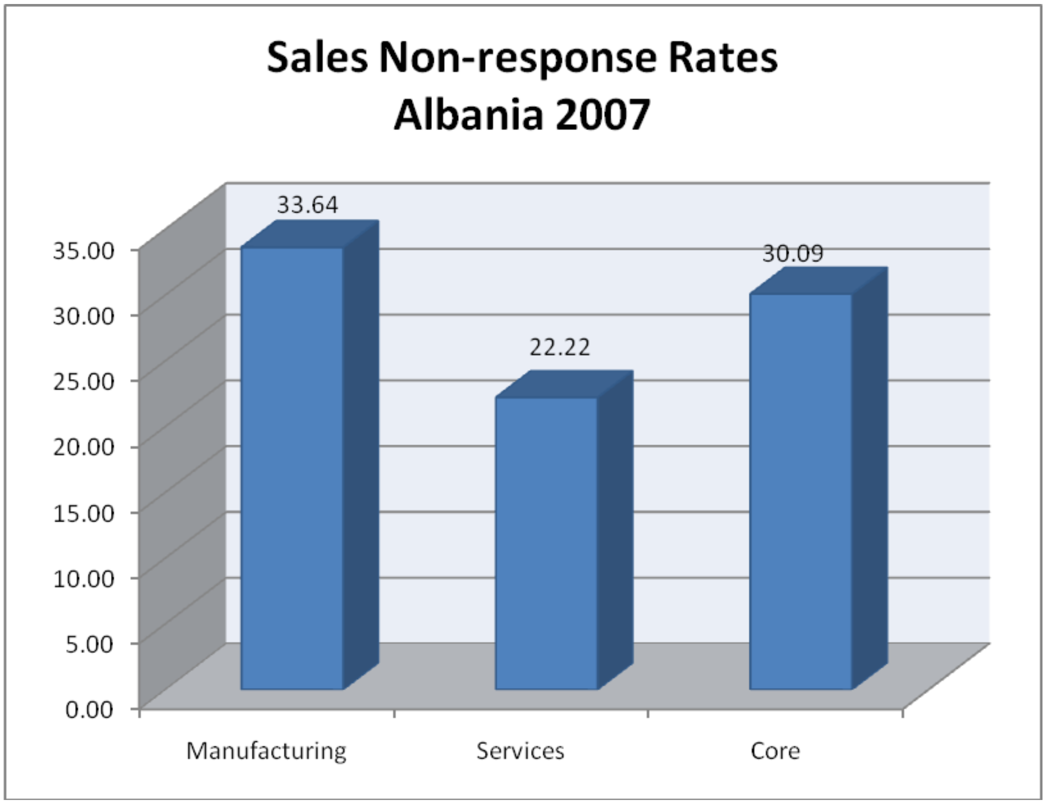
[5] In cases where the number of contacts initially drawn from the sample frame are insufficient to obtain an interview with the targeted number of establishments in a given strata, additional contacts for that strata may be drawn from the sampling frame. If all establishments in that strata have already been contacted and the sample target has not been reached, the sample design may be adjusted to allow additional interviews in other strata.

interview at different times/days of the week before a replacement establishment was allowed to be contacted for an interview.

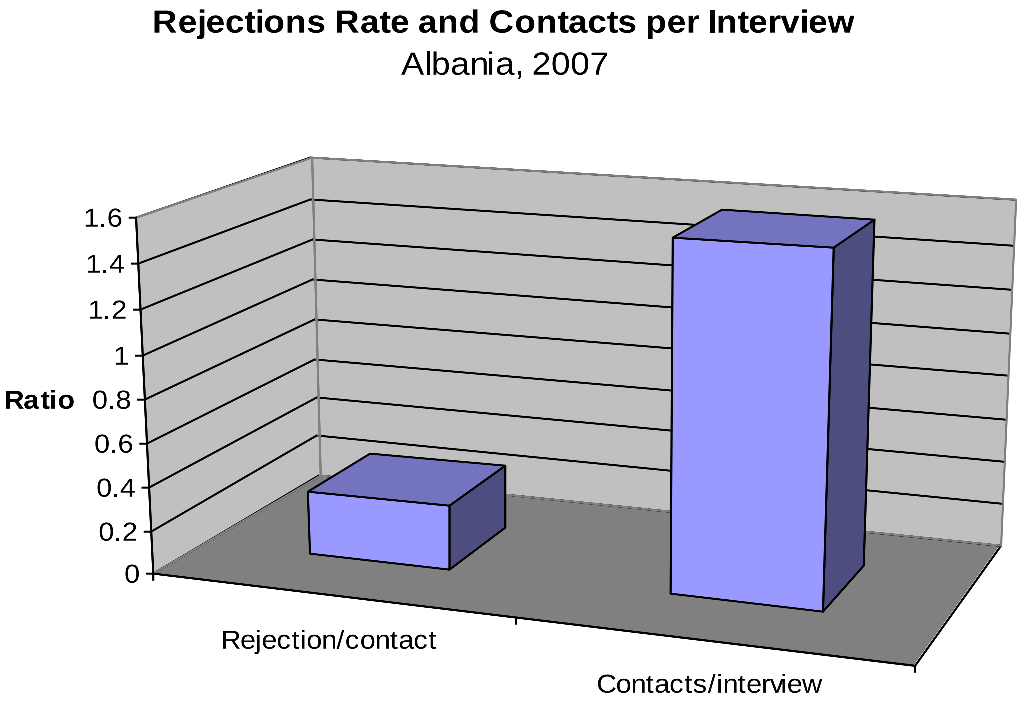
Further research is needed on survey non-response in the Enterprise Surveys regarding the potential introduction of bias through substitution and non-response.

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 as a different option from don't know (-7).

Special attention was paid to the variables needed to assess performance at the establishment level. Establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of low response. The following graph shows non-response rates for the sales variable, *d2*, by type of questionnaire.



As the following graph shows, the number of contacted establishments per realized interview was 1.48. 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 (e.g., establishments that closed or were in ineligible sectors). The relatively low ratio of contacted establishments per realized interview suggests that the main source of error in estimates in the Republic of Albania may be selection bias and not frame inaccuracy. Refusal rates are also shown in the graph below. For each establishment eligible for an interview, 0.18 refused to participate.



Details on rejections rates, eligibility rates, 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 the Republic of Albania. All enterprise surveys suffer from these shortcomings but in very few cases they have been made explicit.

References

Cochran, William G., Sampling Techniques, 1977.

Deaton, Angus, The Analysis of Household Surveys, 1998.

Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.

Lohr, Sharon L. Samping: Design and Techniques, 1999.

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

Appendix A

Cell weights

Albania

	Classification		
SIZE (employee number)	Manufacturers	Services	Residual
4-20	11	10	6
21-100	8	14	7
101+	16		12
NA	59	51	22

Note: BEEPS & Large Tax Payers Samples have weights of 1.

Appendix B

Status Codes

Codes	Albania
1. Eligible establishment (Correct name and address)	305
2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	0
3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	0
4. Eligible establishment (Wrong address - the firm/establishment has changed address and the address could be found)	6
5. The establishment has less than 5 employees	0
6. The firm discontinued businesses/ unattainable	1
7. Not a business: Private	0
8. Not a business: Education or Government	0
9. No reply (after having called in different days of the week and in different business hours) out of order, no tone	42
10. Answering machine	0
11. Fax line	0
12. Wrong address/ moved away and could not get the new references	3
13. Refuses to answer the screener	174
14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0
15. Out of target - cooperative, outside the covered regions	285
	816

Response Outcomes

Outcomes	Albania
1. Complete effective interviews	304
2. Incomplete effective interviews	0
3. Refusal	174
4. In process to make an appointment (they have already answered the screener)	0

Appendix C

Eligibility Rules

Status Code	Eligibility Criteria		
	Strict	Weak	Median
1. Eligible establishment (Correct name and address)	1	1	1
2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	1	1	1
3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	1	1	1
4. Eligible establishment (Wrong address - the firm/establishment has changed address and the address could be found)	1	1	1
5. The establishment has less than 5 employees	0	0	0
6. The firm discontinued businesses/ unattainable	0	0	0
7. Not a business: Private	0	0	0
8. Not a business: Education or Government	0	0	0
9. No reply (after having called in different days of the week and in different business hours) out of order, no tone	0	0	1
10. Answering machine	0	1	1
11. Fax line	0	1	1
12. Wrong address/ moved away and could not get the new references	0	0	1
13. Refuses to answer the screener	0	1	1
14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0	0	0
15. Out of target - cooperative, outside the covered regions	0	0	0

Establishments Estimates

	Eligibility Criteria		
	Strict	Weak	Median
Albania	2588		

Appendix D

Questionnaires:

Problems for the understanding of questions (write question number)	No major problems
Problems found in the navigability of – questionnaires (for example, skip patterns).	No major problems
Comments on questionnaires length:	The length of the questionnaires was a problem because businessmen are usually very busy. Sometimes many got "frightened" just by looking at the number of pages of the questionnaire, and refused to participate in the survey.
Suggestions or other comments on the questionnaire:	Questionnaires should be shorter, in order to improve the response rate and to keep a high quality of the interviews.

Database

Comments on the data map	None
Comments on the data processing	None

Country situation

General aspects of economic, political or social situation of the country that could affect the results of the survey:	Because a big part of the Albanian economy is informal, businesses usually keep two financial books. One for their own purposes and the other one for the tax office. Because of this, for some cases we are not if firms provided us with their real figures or the reported figures. The same issue affects also the declaration of employees.
Relevant country events occurred during fieldwork:	The fieldwork started in difficult period of the year, just ten days before Christmas and New Years eve. This period of the year is a closing period for business activity and affected the response rate that we achieved. For this reason, fieldwork end was delayed.
Other aspects:	None