Firms and informality a new data base for Colombia

Cristina Fernández

Documento de Trabajo
Alianza EFI - Colombia Científica
Noviembre 2021

Número de serie: WP5-2021-009

FIRMS & INFORMALITY
A NEW DATA BASE FOR COLOMBIA

Cristina Fernández[1]

Abstract

The article Firms & Informality identifies the main stylized facts of business informality in Colombia. To identify these facts, an innovative database (EEG) was built by attaching information from the Micro-business Survey (EMICRON), the structural surveys of manufacturing, trade, and services (EAM, EAC and EAS) and the questions asked to the employer by the household survey. Additionally, some missing information was completed by using the GEIH questions addressed to the workers. According to the exercise the main features that a model of informality in Colombia should consider are the following: 1. The share of microbusiness and self-employment is high. 2. Although informality is decreasing in firm´s size it might not be assumed as restricted to microbusiness. 3. The intensive and extensive margin of informality should be included independently, because they do not always move in the same pace/direction 4. Both worker´s and firm´s heterogeneity should be considered since the reasons to be informal vary among informality types, but informality might also cause further heterogeneity. The heterogeneity variable should be observable to allow the implementation of policy recommendations. 5. Some institutional features that can be contemplated in a model are: the large, exempted tax brackets, the profit taxes formal labor cost deduction, and the limited scope of the single tax scheme that causes a rather continuous than binary variable of informality in the country.

JEL classification: J46, O17, L11, O47

Keywords: Informality, Firm informality, Informal labor market, Taxonomy of informality, Policy recommendations for informality, Size distribution of firms

[1] Economics PhD student at Universidad del Rosario. I would like to thank the unvaluable help of Andres García-Suaza in developing this paper. Juan Miguel Gallego, Fernando Jaramillo, and Alain Desdoigts were also of great help, not only through our weekly meeting, but also reading, hearing, and commenting multiple times the same story. Marlon Salazar and Camilo Rios were very helpful in the initial steps of this paper, and Monica Ortiz provided useful comments to the paper. This working paper is funded by the Colombia Científica-Alianza EFI Research Program, with code 60185 and contract number FP44842-220-2018, funded by The World Bank through the call Scientific Ecosystems, managed by the Colombian Ministry of Science, Technology and Innovation. I also would also like to thank el Departamento Nacional de Paneación (DNP), that financed a precious version of this paper.

I. INTRODUCTION

After many years of studying labor and firm informality someone asked me the following question: which stylized facts do you consider that an informality model for Colombia should include? I soon realized that although I have several answers in my mind, these pieces of information could not be merged because I didn't have a comprehensive database encompassing not only the whole economy but also the workers and the firm perspective.

I therefore decided to create a new database, by holding together several available surveys, that allowed me the identification of the following stylized facts: 1. Firm informality in Colombia is not a binary condition; 2. Labor and firm informality, although decreasing in firm-size, are present through most of size and sector distribution of the economy; nevertheless, there are some subsistence firms and workers with few possibilities to be integrated into the formal economy; 3. Both labor and firm informality heterogeneity are important to be considered in a model, but if one needs to be chosen I would go for the latter, otherwise it means ignoring the vast amount of informal, unproductive and small amount of firms that comprise a large portion of the Colombian economy. Firm and labor heterogeneity implies heterogeneity of policy recommendations.

Self-employment was not left out of the discussion because it is not important but, on the contrary, because it is too important. Colombia has a self-employment rate that almost doubles the one of the other countries of Latin America and there are some signs that suggests that this kind of employment is even more segmented, and not always obey to the traditional cost-benefit analysis but considers other features as independence and flexibility (Fernández and Mejía, 2020). I, therefore, believe that it is important to understand firm dynamics first, to only include later self-employment.

The importance of this analysis lies on the fact that Colombia does not have a recent Economic Census; being the last one performed in 1991. This situation contrasts with the strength of sectoral surveys, and particularly the strength of the EAM (Encuesta Anual Manufacturera), the EAC (Encuesta Annual de Comercio) and the EAS (Encuesta Anual de Servicios), and of the Household Survey in Colombia, that has a special module of micro firms (EMICRON) and an important set of questions directed to entrepreneurs as firm-size, income, registration to the authorities and the reasons to be independent, among others. It is possible to argue that the number of entrepreneurs in the GEIH is representative of the number of firms in Colombia, by assuming that each entrepreneur represents one firm and given that the number of entrepreneurs is representative of the total number of entrepreneurs, at quarterly frequency.

Under this perspective, this paper constructs a database by holding together the EMICRON, the EAM, EAC and EAS; and filling the with the GEIH questions oriented to entrepreneurs. The set of variables collected through this procedure is rather wide and allows me to generate a good characterization of firm informality in Colombia. However, there were some gaps in identifying the workers characteristics for bigger firms, that are fulfilled through estimations using the GEIH's workers module and the EMICRON.

This effort goes in line with the recommendations made by the government document CONPES (DNP,2019), which calls for the need of strengthened data bases to understand informality and firms' behavior. This new database can be understood as a base material for the expected 2022 Economic Census. Also, in this path, Dane has created recently the Directorio Estadístico Empresarial (DEE), which uses the administrative social security data (PILA), the Chamber of Commerce registration data base as well as the structural survey (EAC, EAM, EAS) to create a directory of firms. However, the microdata collected through this effort is not available to the public and the set of variables collected is much smaller. Therefore, I consider that these two exercises complement each other. On top of allowing Dane to find which pieces of information are missing, this exercise can bring a periodic and publicly available description of the firm's universe and informality, provided a good fit between the collected data base and the expected Census.

Another feature of the collected database is that most of it is attached to the household survey. This means that it allows the communication of two areas of knowledge that traditionally follow different paths: productivity and welfare economics. Therefore, it allows to analyze the impacts of productivity policies over welfare, and the impact of the characteristics and welfare of entrepreneurs and workers over productivity; being informality a key channel between the two avenues. This can be a future extension of this paper.

However, this database is not exempted of limitations, being the most important, that the household weights were not created with the description of the firm's universe in mind. Particularly, it might be the case that I am not getting firms outside the structural surveys that are not owned by an entrepreneur but by a different kind of ownership. Nevertheless, population weights mostly obey to availability of information, which suits the needs of creating a demography of firms. It is also possible that, despite the efforts made by Dane to capture relatively small firms in the structural survey, there remains some smaller firms to be included. To identify these gaps, I perform some consistency checks.

This paper is structured in six sections, of which this introduction is the first one; the second section contains a literature review; the third section describes the methodology used to construct the data base; fourth and fifth sections presents the stylized facts of informality at firm and at worker level, respectively. Conclusions and research agenda are resumed in section sixth.

II. LITERATURE REVIEW

Perry et al (2007) made one of the first attempts to understand the relationship of the firm dynamics and informality at the Latin American Level. According to their findings, the informal firms encompass not only small subsistence firms but also firms of a larger size that fail to comply with the regulations. They also suggest that considerable efficiency gains can be obtained through moving resources from low productivity firms to high productivity firms. Concerning policy recommendations, they argue that some firms can benefit of lowering the costs of informality and react to an increase in the costs of being informal, but they also understand that

the best policy for small firms require policies as access to formal credit, training, and business development services.

One of the most exhaustive exercises to describe the taxonomy of firms in the context of informality was performed by Levy (2018) for the case of Mexico. According to Levy the cause of low productivity in Mexico is the misallocation of resources, very much linked to informality, and non-salaried work. To arrive to this conclusion, he performed an exhaustive firm-level analysis, identifying several styled facts related to informal Mexican firms; 1. They absorb a significant amount of capital and labor. 2. They can be found through all the sectors of the economy, and al the territory. 3. They are not necessarily illegal because most of them hiring workers through non-salaried work (that is legal in Mexico), and not necessarily completely informal (there is a scale of informality). 3. Most informal firms are very small (lees than 5 workers), but not all informal firms are small. The proliferation of small firms generates a proliferation of entrepreneurs over workers, that is not always optimal. 4. The composition of economic activity shifted towards the informal sector over time. 5. Firms that hire formal workers are the most productive, and in general formal firms are more productive but, not every formal firm is more productive than each informal firm.

This analysis leads Levy to the premise that low productivity firms absorbed more capital and labor than they should, while more productive ones fail to receive sufficient resources; and this process is reinforced by the longevity of those small unproductive firms. Furthermore, the small unproductive firms behave in a way that is socially inefficient. For example, choosing to be informal when they can afford to be formal. By being informal they must be far from the eye of the authorities and therefore, do not accept cash, do not grow, do not make use of technology, and do not hire formal workers. On some cases they have high entry rates and low survival rates, and therefore generate short lived jobs. This analysis was summarized later by Alvarez and Ruane (2019) to estimate Ulyssea (2018) for the case of Mexico.
More condensed and equally rich, is the description of firm informality performed in Ulyssea (2018), Ulyssea (2019) and Ulyssea (2020). According to them, informal firms are on average smaller, run by less educated individuals, pay lower wages and are less productive than formal firms. He also finds that despite this differences, formal and informal firms in Brazil coexist across sectors and levels of productivity, and they also argue that there is no evidence of the missing middle-size firms characteristic of a dualist economy. Finally, he finds that the wage gap between formal and informal workers -characteristic of segmented economies since it means that formal and informal workers perform different task in the economy- disappears when he controls for firms' characteristics. It also means that self-selection is one of the drivers of the wage gap. Another stylized fact found by the author are decreasing levels of the intensive and extensive margin of informality on firms' size, indicating that the cost of operating informally is increasing. The author also finds that the process of dynamic selection takes place in both formal and informal sectors, but is weaker in the later, and therefore the ex-ante heterogeneity rather than ex-post is the determinant of firm's dynamics.

Based on all this stylized facts Ulyssea developed a model where the less productive firms cannot even pay the fix costs of being formal; the medium productivity firm might pay those fix costs of

being formal but find that is more profitable to be informal, and the more productive firms, that are a minority but also the biggest, are the only ones that find profitable to be formal. This perspective has the advantage of integrating in one model two points of view of informality that until now were considered as excluding: the dual market perspective (Lewis, 1954 and Harris y Todaro,1970); and the De Soto margin according to which informality is caused by high entrance costs (De Soto et al., 1989 y De Soto 2000). Other advantage of Ulyssea (2018) is that is a general equilibrium model that takes into consideration both the intensive and the extensive margin of informality.

For the case of Colombia, Eslava, Haltiwanger and Pinzon (2019) analyzed the entire population of non-micro manufacturing establishments in Colombia and found that the size distribution of manufacturing formal firms in Colombia exhibits a high concentration of small old firms, in terms of number, and employment, pointing to a shortage of high-growth entrepreneurship and a relatively high likelihood of long-run survival for small, likely unproductive firms. The authors consider that this is a key feature of Colombian low rates of productivity growth. More focused on informality, Fernandez (2019) made a first attempt to understand firm informality dynamics but based on the Encuesta de Microestablecimientos (2013-2016), which covered only microbusiness and was not a representative survey.

Although these papers made important contributions to the understanding on firm's dynamics and informality in Colombia, they relied on partial sets of information, and therefore, cannot address some of firm informality features identified for Mexico or Brazil. Therefore, I collected several surveys in the economy to create a new database that can help to address these issues, which methodology is described in the next section. It is only a first step, since key surveys as EMICRON are new; and therefore, the questions related to dynamics cannot yet be resolved.

III. METHODOLOGY

a. Sources of information:

The data bases used to conduct the empirical analysis are the following:

- **EMICRON**, a microbusiness survey that not only is representative of the firm population, but also provides new information for estimating the value added, approximating different measures of informality, and analyzing some characteristics of the entrepreneurs as their level of education. EMICRON excludes, government business, airplane trips, the financial and insurance sector, and some areas (Dane, 2020d).[2] Self-employment was also excluded.

[2] It includes microfirms in 24 cities and their respective metropolitan areas, but does not include Arauca (81), Casanare (85), Vichada, Guainía (94), Guaviare (95), Vaupés (97), Putumayo (86) y Amazonas (91). San Andres (88) data is excluded since it is not included in GEIH (alternatively we can include it in GEIH). The agricultural sector is not excluded, as in Ulyssea. The original data base of Emicron 2019's has 101 681 workers' observations that using weights expands to 6'894 711 workers and 86 969 firm's information that expand to 5'874 177 firms. Excluding San

- **Gran Encuesta Integrada de Hogares (GEIH) worker's level (dec 2018 – Nov 2019).** The household survey in Colombia is the source of the business interviewed by the EMICRON survey (Dane, 2013). However, the dates are different than EMICRON's since the inclusion in EMICRON is selected one month in advance.[3]

- **Gran Encuesta Integrada de GEIH firm level (dec 2018 – nov 2019):** The household survey has a set of special firm questions performed over independent workers (Dane, 2013). From them is possible to create a database of firms, assuming that each entrepreneur has only one firm and that the number of entrepreneurs is representative of the total number of entrepreneurs. Second entrepreneurs' jobs are not considered, because the variables describing the job are different and usually not available. As in the rest of the exercise, self-employment and workers that report to work alone but are not self-employed (742 observations), were excluded, as well as sectors not covered by EMICRON.

- **Encuesta anual manufacturera (EAM, 2019).** This is a structural and exhaustive survey that considers all manufacturing establishments with 10 or more than 10 employed individuals at national level. The information corresponds to 2019 but was collected in 2020. During 2019, 7631 establishment of which 6742 have 11 or more workers, involving 706.563 occupied population (Dane, 2020b)

- **Encuesta anual de comercio (EAC, 2019).** This is a structural and exhaustive survey that considers all retail economic units with 10 or more than 10 employed individuals at national level, exempting used good retails (cars resales are included) (Dane, 2020a). The information corresponds to 2019 but was collected in 2020. During 2019, 9859 economic units of which 8033 have 11 or more workers, involving 632.270 occupied population (remunerated and non-remunerated workers)

- **Encuesta anual de servicios (EAS, 2019).** This is a structural and exhaustive survey that considers all service formal economic units at national level, exempting used good retails (cars resales are included). In contrast with EAC and EAM, the cut level differs with the subsectors. A detailed list of cut levels by sectors can be found in Dane (2020c)[4]. The information corresponds to 2019 but was collected in 2020. During 2019, 6398 economic units were interviewed, of which 5.985 have information about the number of workers and

---

Andrés (1434 observations) there are 100247 workers' observations and 85756 firm's observations, of which 75.122 are self-employment and were excluded.

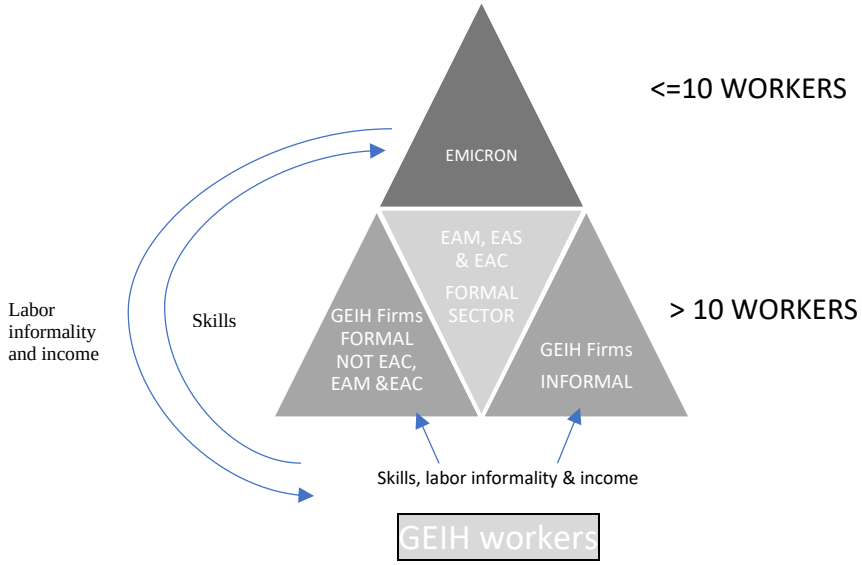
[3] Initially the GEIH has 337,507 observations, representing 22'303,304 workers (average for the whole year, weights/12). However, it's needed to exclude some sectors to make it comparable to EMICRON: The following sectors were excluded: government (12743), utilities (3447), air transportation (230), financial and insurance sectors (4815), and household activities acting as employers (12002). Additionally, we merged both data bases to obtain the following data that was lacking in EMICRON: imputed skill level of workers in EMICRON, using discriminatory analysis techniques (detailed in the next section), school level and hours worked of entrepreneurs and four-digit CIIU level (according to the entrepreneur answer).

[4] https://www.dane.gov.co/index.php/estadisticas-por-tema/servicios/encuesta-anual-de-servicios-eas

less than 10 workers involving 2'054.164 occupied population (remunerated and non-remunerated workers).

The universe of firms in this paper is constructed in the following way: firms with less than 10 workers are represented by EMICRON; firms with more than 10 workers (11+) are represented by the structural surveys (EAM, EAS and EAC) after discarding firms with less than 10 workers in those surveys and by the GEIH if they are informal or belong to a sector or a size not covered by the structural surveys. The new data base from here on is going to be called EEG (EMICRON-Estructurales-GEIH). Diagram 1 shows the sources of information of EEG.

Diagram 1. Sources of information of the EEG



The observational unit used in this database is the economic unit that comprises firms and establishments hiring at least one paid worker, meaning that self-employment is left out of the accounts[5]. In a second phase of this project, it should be possible to include self-employment as an additional sector. This observational unit is comparable to the hiring units suggested by OECD manual (2007). The size of the firm in terms of workers includes direct workers, direct and indirect contractors, interns, non-remunerated workers and partners and entrepreneurs[6]. However, sometimes the aggregate "*dependents"* is used. Dependents includes salaried workers, partners, and no waged workers, but exclude the entrepreneur.

Table 1 summarizes the data used in this document. There are using 31,163 observations of which 10739 come from EMICRON, 6792 from EAM, 8033 from EAC, 5990 from EAS and 590 from GEIH These observations represent 793,359 firms, of which EMICRON accounts for the 93% and GEIH

[5] Firms with only one worker (that can be the boss) are also excluded from the analysis
[6] The inclusion of contractors and interns is important to measure productivity, however, if self-employment is included, we can have double accounting, as they might appear as self-employers.

for 5%; and 6,5 million workers whose information is collected from EMICRON (35%), GEIH (13%), EAM (11%), EAC (10%) and EAS (32%).

Table 1. Composition of EGG database

	EMICRON	EAM	EAC	EAS	GEIH	Total
Firms (# observations)	10.739	6742	8033	5990	590	31.163
Firms (population)	10.739	6742	8033	5990	590	32.094
Share in the total # of firms	735.591	6742	8033	5990	37.003	793.359
Workers (# observations)	93%	1%	1%	1%	5%	1,01
# workers (average)	10.739	6742	8033	5855	590	31.959
# workers (total)	3	105	79	351	22	8,2
Share in the total # of workers	2.296.557	706.563	632.270	2054164	824.860	6.514.414

Fuente: EEG

How does this compare to the universe of firms? According to the preliminary data of the 2022 census there are 2.5 million economic units in Colombia, but this includes self-employment. According to the GEIH the amount of self-employed that own a business are 1,6 million, leaving us with 900 economic units bigger than 10 workers. Similarly, as shown in Table 2 the DEE finds 5 million of economic units, of which, approximately 3.5 million are self-employed (maybe more since I am not considering firms that do not report number of workers), leaving 1,5 million of firms bigger than 2 workers. However, the 6,5 million workers that we are covering are far from the 10 million workers that can be identified in the GEIH (covered by EMICRON's sectors and excluding self-employment) and in the DEE. Nevertheless, is important to have in mind that the DEE only records economic units recently registered in either RUES/RUT or PILA, and therefore the number of units can be even bigger than the record.[7] Another piece of information, are the formal entrepreneurs that hire formal workers (429.000), accounted by the administrative records, and processed by RELAB (Dane); however, the comparison of the databases depends on the informality criteria that can create a number of formal firms that ranges from 77 to 463 thousand

Table 2. DEE (DANE) and own estimations

	Total	**1 SS register**	**Firms without SE**
Mono activity	2,7	1,8	0,9
Multi activity	2,3	1,7 (e)	0,6
Total	5,0	3,5	1,6

Source: Dane and own calculations

On top of this external validation, it is possible to perform an internal validation of the data by comparing the segments of the firm population that overlap. This cross validation is presented in Annex 1.

[7] It is also true that administrative records can capture firms only created for tax reasons.

b. Definitions

The main definitions for the purposes of this paper are firm informality and labor informality. All other definitions can be found on Annex 2.

- **Firm informality (extensive margin)**: This paper understands of firm informality as continuous rather than a binary function, following the perspective of the government document CONPES (2019). However, to facilitate the analysis, we only consider three scenarios of firm informality:

    1. *Strict informality*
        a. Renewed or new registration to the Chamber of Commerce (or RUT in GEIH).
        b. Formal accounts (General Balance or Profit and Losses Statement or with a daily register book)
        c. Pay taxes, are not supposed to pay taxes (micro firms), or have more than 10 workers (GEIH)
    2. *Consistent informality*
        a. Renewed or new registration to the Chamber of Commerce or RUT.
        b. Formal accounts (General Balance or Profit and Losses Statement or with a daily register book)
    3. *Relaxed informality*
        a. Renewed/new Chamber of Commerce registration (or RUT for GEIH)

It should be noted that the register of chamber of commerce and formal accounts are only used as proxies of formality since some regulatory exemptions apply[8]. Through this paper the mostly used definition of informality is the strict informality, since it resembles more closely the formal and informal firms in theoretical models. Firms whose information comes from EAC, EAS and EAM are assumed formal in under any of the 3 definitions of informality.

- **Labor informality (intensive margin):** The percentage of workers hired informally is known as the firm's intensive margin. I identified as informal those workers that neither make health or pension contributions. Non remunerated workers, partners and entrepreneurs are also classified according to their informality status and therefore, included in the denominator[9]. All workers in firms reported by EAC, EAM and EAS are assumed to be formal.

c. Estimation of missing information in EEG

One of the advantages of EMICRON is that it is an employee-employer survey that allows to measure simultaneously the intensive and extensive margin of informality, as well as firm's and

8 DEE uses as an alternative to these criteria the affiliation to SAYCO and ACINPRO.
9 In the GEIH, informal workers were identified as those that do not contribute to pension or did not answer the question and are not registered or appear as beneficiaries of the contributory health system or are registered in the subsided regime despite having a job. In the EMICRON, informal workers are those that do not contribute to pension and/or do not contribute to health.

worker's income. Whereas the GEIH, lets to identify labor and firm informality and labor income and firms profits but does not allow the interaction of workers and firms statistics. However, there are some common variables between the questions asked to workers and the questions asked to workers such as the range of the size of the firm, its location, and the population weights[10] that makes possible the estimation of some variables at a firm level, using workers level information. Likewise, it is possible to use the information contained in EMICRON, that has both the workers and firm's perspective, and again some common variables to complete these estimations.[11] Diagram 1, shows the flow of information between surveys.

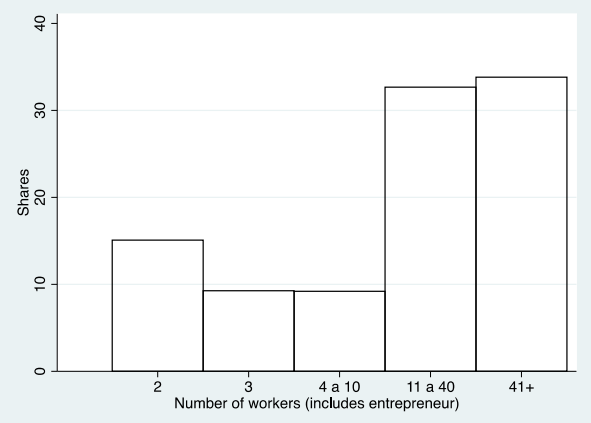
IV. FIRMS AND FIRMS INFORMALITY FACTS

a. Size

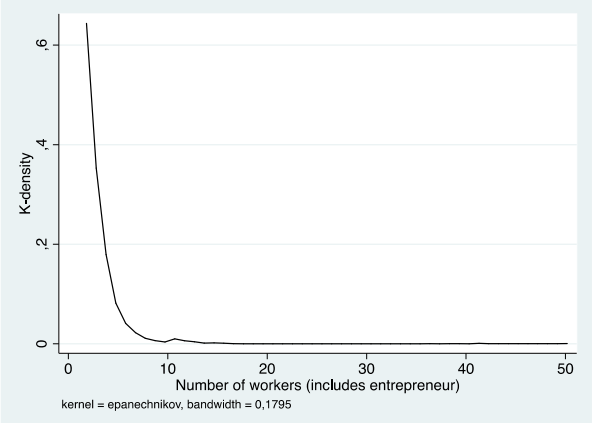
Figure 1A presents the histogram of firms by size in terms of number of workers. As it is shown data availability is higher for relatively bigger firms. Figure 1B, shows the same information but using weights and softened using a kernel. This figure confirms that Colombian production structure is made of very small firms. Excluding self-employment 93% of economic units are microbusiness and according to the DEE that includes self-employment, this percentage is 98%.

Figures 1A y 1B. Firms by size

1A. Histogram



1B. K -density



Source: EEG. the minimum value is 2, because our firm definition only considers workers with a remunerated dependent worker.

One of the advantages of the EEG is that it allows to validate some assumptions used by the informality literature as the shape and the parameters used to describe the universe of firms.

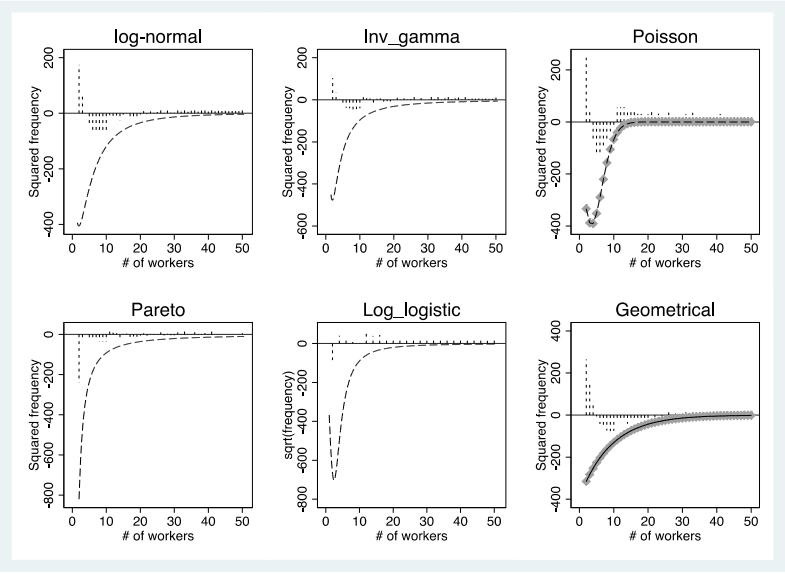
[10] Weights collect relevant information not observable in the surveys

11 In practice, to mix these two sets of information (EMICRON and GEIH workers), an estimate of the variables of interest was made using information from the GEIH and this estimate was used as a control variable in a second estimate, using the information from the EMICRON. Although not used in this analysis, the level of qualification is an important variable to add some workers heterogeneity to the analysis. However, it is not collected on EMICRON and neither in GEIH (firms) or structural surveys (EAC, EAM and EAS). To overcome this problem, I estimated an inverse of the Mincer equation on GEIH and used the coefficients in the surveys missing the skill variable. Skills in EAM were estimated by using the distinction between operators, administrative personnel and technicians and professionals.

Figure 2 shows the squared errors incurred in estimating firm's number of workers distribution through several distributions. The most used distributions by the literature are the log-normal and the Pareto. However, as shown in the graph, the inverse gamma and the log-logistic appear as alternative good options[12].

Although the Pareto approximation generates fewer overall errors, it is too inaccurate to describe firms with two workers. Probably the most realistic function is a Pareto-log-normal distribution, where the shape of the distribution is that of the log-normal function, but the tail of the distribution takes the form of a Pareto distribution[13]. Additionally, it is possible that a more exact fit between the theoretical approximation and the data may be achieved if self-employed businesses are included.

Figure 2. Firms by number of workers approximated using theoretical distributions. Firms with less than 50 workers.

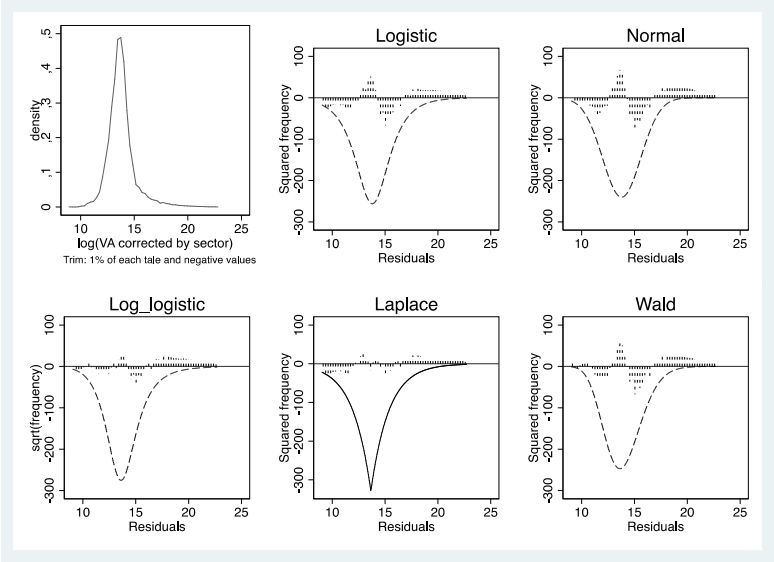


[12] The lognormal distribution was approximated with a location parameter of 1.2 and a shape parameter of 0.74 according to Jenkins and Van Kerm (2007). The Pareto distribution was approximated with a location parameter of 7 and a minimun size of 2 a shape parameter of 1.7 according to Jenkins (2007). The parameter used in the inverse gamma distribution were 3,17 and 9,02 according to Cox and Jenkins (2011).
[13] As suggested by Ulyssea (2018) for the distribution of signatures according to their added value

Fuente: EGG using Buis (2007), Jenkins (2007) and Jenkins and Van Kerm (2007). Data was collapsed by the number of workers in order convert pweights in frequency weights, that are the only ones allowed by the program.

Size can also be measured in terms of value-added. Figure 3 shows the value-added distribution, and the squared errors incurred in estimating firm's value-added distribution through several distributions. According to the graph, the log-logistic and the Laplace distribution are those that fit the data better. Annex 1 contains the details of how the value-added is collected and estimated, in some cases.

Figure 3. Firm by value-added approximated using theoretical distributions.



Fuente: EGG using Jenkins (2007)). Data was collapsed by the number of workers in order convert probability weights in frequency weights, that are the only ones allowed by the program.

b. Extensive margin

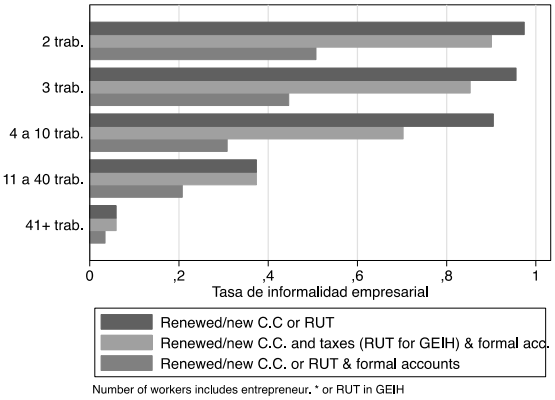
Figure 4A shows firm informality rates according to the three measures of informality used in this paper: the Strict definition, the consistent definition, and the flexible definition. As it is shown in the graph, the informality rate is widely spread in small firms even if self-employment excluded, suggesting a more structural problem. Informality is also rather high for firms with more than 10 workers and therefore assuming that these firms are formal (as in Ulyssea, 2018) is not realistic for the Colombian case. There are even some firms with more than 50 workers that are informal under the strict informality definition, but I didn't find any characteristic that allowed me as flaws of the informality definition.

Figure 4B shows firm informality for microbusiness for Colombia and Brazil. According to this graph, small Brazilian firms show informality levels alike Colombia' stricter informality scenarios, but firm informality in Brazil decreases much faster on firm's size. The information derived from both graphs suggests that whereas in Brazil it is a plausible assumption that firms bigger than 10 workers are formal, this is not a valid assumption for the Colombian case. Another feature that should be considered is that the existence of the single tax scheme comprehensive of social security in Brazil makes it easier to understand this feature as a binary variable, whereas in the

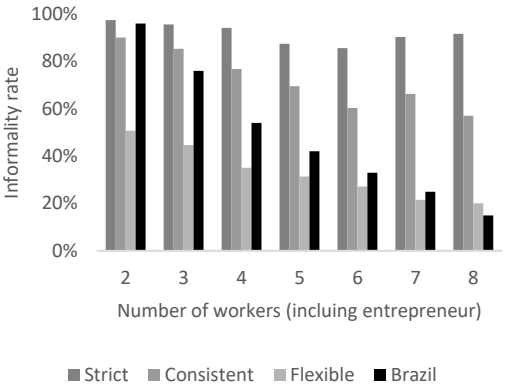
case of Colombia, as explained before, it is more of a continuous nature, and at least several scenarios should be considered in the analysis.

Figure 4. Extensive margin of informality by criteria

4A. Colombia by ranges



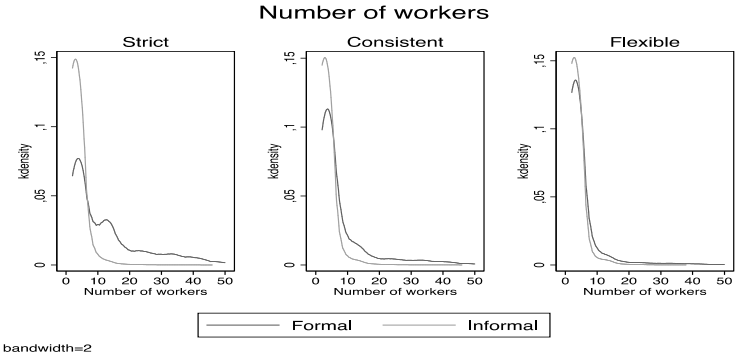
4B. Colombia and Brazil, microbusiness

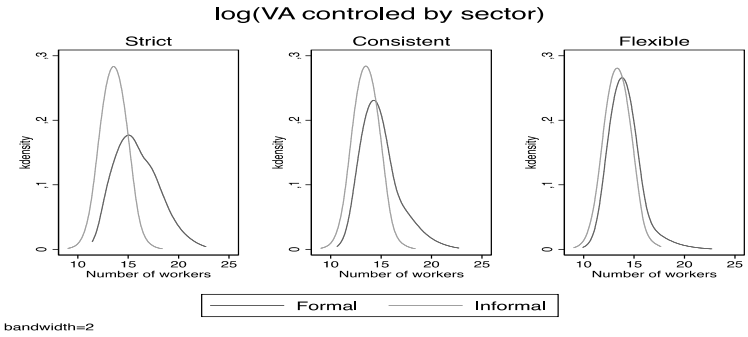


Source: EGG and Ulyssea (2018)

The behavior of the extensive margin suggests that the distribution that better describes firm structure might be different for formal and informal firms. However, the distribution of firms by size measure and informality criteria in Figure 5, suggests that this is only the case when the stricter criterion of informality is used, and might be related to the amount of data available.

Figure 5. Distribution of firms by size measure and informality criteria





Source: EEG (2019)

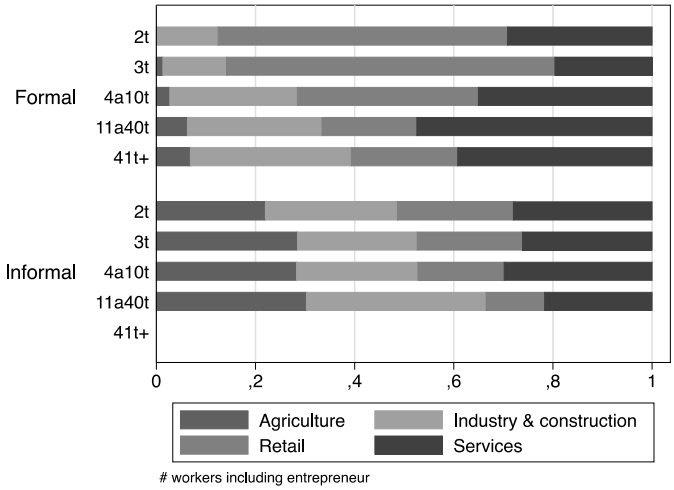
In sum, it is possible to conclude that firm informality in Colombia is higher and more widespread among the number of workers of the firm, than in the Brazilian case. It is also less of a binary condition than in Brazil, perhaps due to the prevalence of the single-tax scheme in the later. This evidence suggests that is not possible to assume that all firms other that microbusiness are formal in Colombia, as in Ulyssea (2018). In turn, this stresses the importance of using a source of information that includes informal firms of relatively higher size as the GEIH.

The evidence also stresses the importance of treating informality as a continuous variable or at least to use different scenarios, being conscious that a more flexible scenario of informality imply less differences between the sizes of formal and informal firms. The modelling and calibration of an informality sector that pays full taxes and contributions needs to consider the stricter informality measure, or otherwise, adjust the level of these taxes and contributions.

c. Dualism

Frequently, the Colombian economy is assumed to be dual, meaning that there are two sectors in the economy, one of those is fully formal, productive, and clustered in some sectors; while the other sector is fully informal, unproductive, and clustered in other sectors. This structure is consistent with a missing middle on firm's distribution across sizes, which was not observable on the Colombian case. Moreover, as observed in Figure 6, informality is evenly spread among aggregates sectors, maybe with the exemption of agriculture. The graph also shows a big share of retail trading among microbusinesses.

Figure 6. Firm informality and economic sector



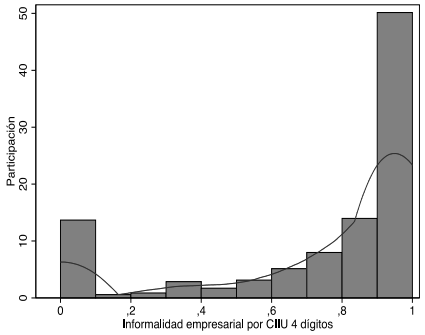
Fuente: EGG

Figures 7A and 7B show the same exercise but at a more desegregated level (CIIU 4 digits) and for the strict and consistent criteria of informality. More specifically it shows the informality rate by subsector organized by quantiles. Therefore, the bar on the right shows the share of subsectors with an average informality rate lower than 10% and the bar the in right shows the share of subsectors with an average informality rate higher than 90%. According to Graph 7A about 50% of the sectors are almost fully informal, which would give us the idea of dualism. However, using a more flexible definition of informality (Figure 7B) the results show a milder dualism in the market. Figures 7C and 7D replicate the exercise for Colombia and Brazil's microbusiness. According to the graph, there is a slighter polarization, and therefore dualism, in the case Colombia.[14]

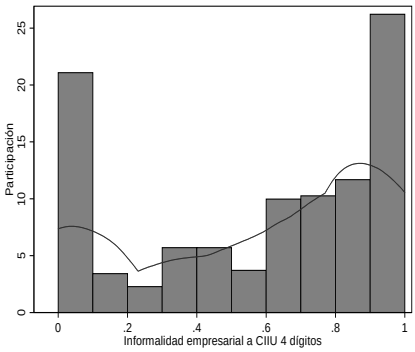
[14] Although Brazil's exercise is performed at a 5 digits level, the number of firms is bigger, and therefore the exercises should be comparable.

Figures 7A and 7B. Firm informality and economic subsectors

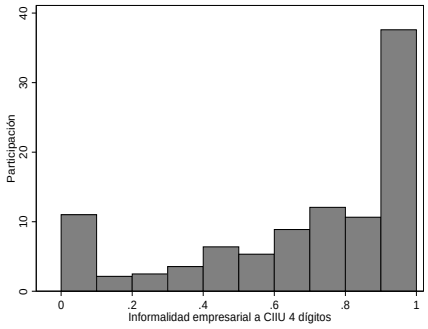
7A. Strict criteria



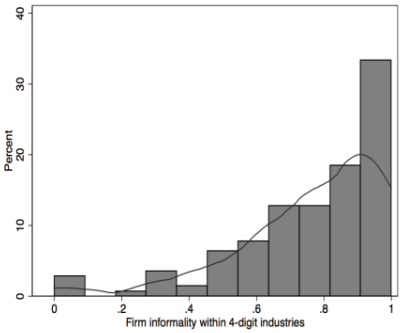
7B. Consistent criteria



7C. Consistent criteria, microbusiness



7D. Brazil, microbusiness



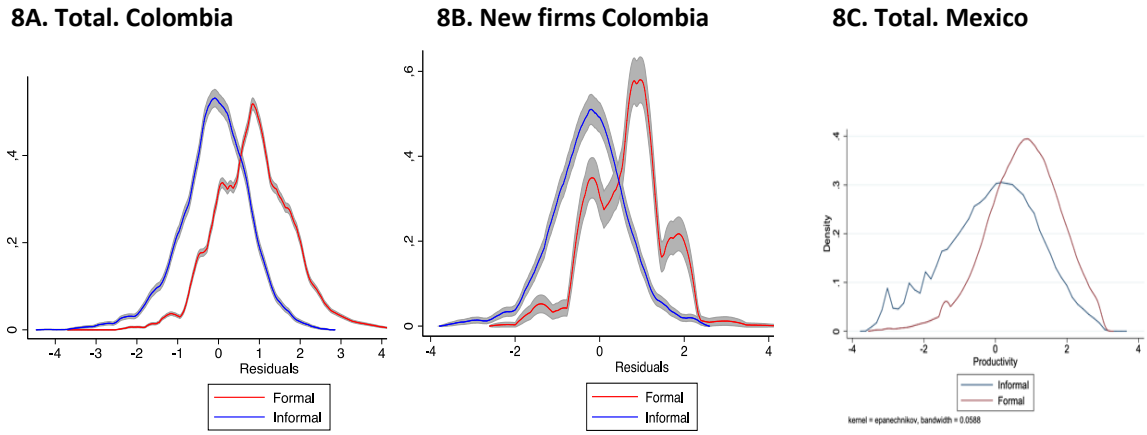
Source: EEG and Ulyssea (2018)

Another sign in dualism is to have an informal sector with very low productivity and a formal sector with rather high productivity. Figures 8A, 8B and 8C show formal and informal labor productivity distributions controlled by sectoral differences and trimmed 1% in each tail of each survey for the whole Colombian survey, newer firms, and the Mexican case. According to Figure 8A, on average, formal firms are more productive that informal ones but there is an important range of productivity shared by both formal and informal firms. The intervals of confidence suggest that this is a significative difference. When comparing the Colombian case (8A) with Mexican (8C), it can be noticed that Colombian informal and formal sector tend to have a more homogeneous behavior within each group (higher density around the mean for the Colombian Case, particularly for the informal sector).

Figures 9A and 9B shows the density of labor productivity controlled by sectoral differences and trimmed 1% in each tail of each survey, for all and new microbusiness. According to Figure 9A, productivity tends to be similar for formal and informal microbusiness, unlike the cases of Brazil and the whole Colombian firm structure. This stress the importance of working with the whole set of firms and not only microbusiness.

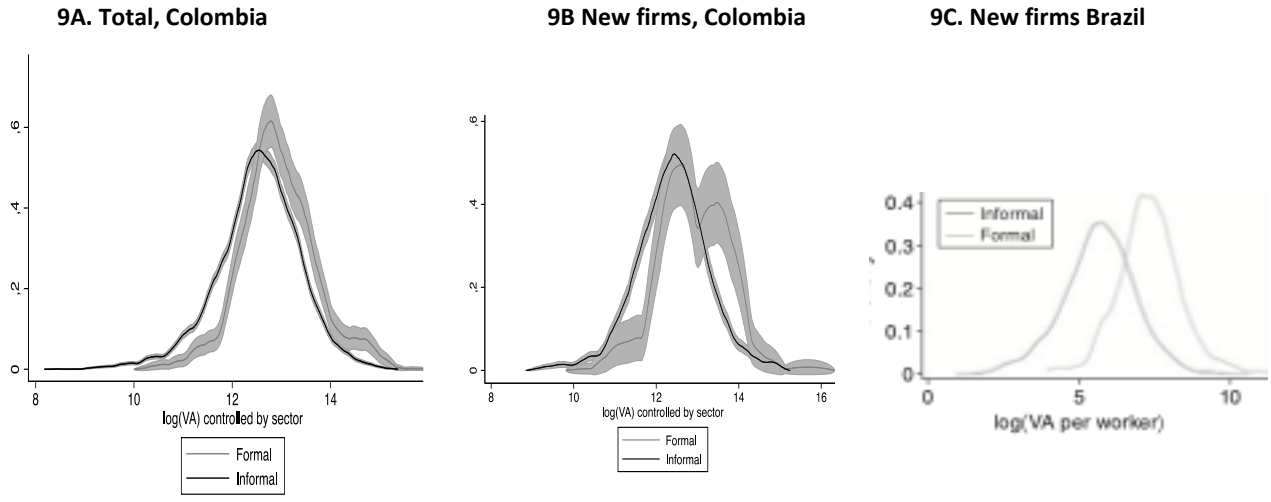
According to Ulyssea (2018), the fact that there are productivity differences between formal and informal new firms, is consistent with firms preselecting into formality or informality before entering the market. The behavior among new firms and microbusiness in Colombia (8B and 9B) does not differ much from the one that includes older firms (8A and 9A), but the intervals of confidence get bigger since of fewer information available. This means that, with the information available is not possible to infer if the heterogeneity of firms is a pre or post market condition.

Figure 8. Value added per worker



Source: EEG (2019) and Alvarez and Ruane (2019). Value added is the residual of the regression of value added and economic sectors at an aggregate level. Constant is not recovered to make it comparable with the Mexican case. The estimation of new firms (less than 2 years) does not include the manufacturing industry because of lack of information.

Figure 9. Value added per worker, microfirms



Source: EEG (2019) and Alvarez and Ruane (2019). Value added is the residual of the regression of value added and economic sectors at an aggregate level plus the constant of the regression. The estimation of new firms does not include the manufacturing industry because of lack of information.

In sum, it is possible to argue that in Colombia there are some signs of both dualism and coexistence of informality and formality in the market; being the dualism clearer when we observe the whole universe of firms. My guess is that the introduction of self-employment would also increase the degree of market dualism in the country. This multiple evidence for both dualism and coexistence of informality and formality in some segments of the market is consistent with the informality models that consider heterogeneity of firms other than their informality status (Perry, 2007 and Ulyssea, 2018) and has important implications for policy recommendations, since the best practices in policies oriented to subsistence firms are very different to those of firms that can actually become formal.

d. Rationality of informality

This section goes further in the argument of preselecting on informality and tries to find the key variable is under the of the entrepreneur's decision of being formal or informal. Ulyssea (2018) argues that there is a non-observable sign of productivity that the entrepreneur gets before entering the market, and according to this sign takes the decision to be formal or not based on a relative profitability criterium.

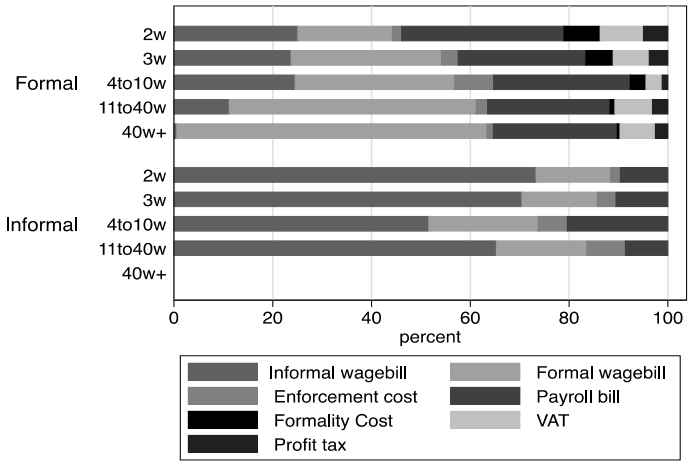
My task in this section is to find an observable proxy to this unobservable variable. Some obvious candidates are the value added per worker for new firms, as suggested by Ulyssea (2018) at the beginning of his paper, and the own characteristics of the entrepreneur as level of education, experience, and his own labor informality condition. Finding this variable is of the most importance, since it is key for implementing and focusing the heterogenous policies to reduce informality that were mentioned before.

To advance in this task, I first estimated the net profits of formal and informal firms and then, I make some exercises to find how it correlates to value added per worker, level of education of the entrepreneur and the reason to create a firm. It is important to stress that this is only a first attempt to find this variable, but further work might involve theoretical modeling. This exercise is performed used the strict definition of informality that is the concept that suits better the condition of a firm that pays all his taxes and contributions.

Firm's costs not included in value added, tend more related to the informality condition. These costs include labor costs, labor and firm's contributions, accounting, associations fees, and taxes. Additionally, informal firms also incur specific costs, determined by the probability of being controlled and punished by the authorities. Figure 10, that segregates those additional costs by size and informality condition, show smaller firms facing higher taxes and formality contributions, an intensive margin decreasing on firms' size and informal firms incurring in some formal hiring. This figure also includes the cost of enforcement that tends to be increasing in informal firm's size.[15] The details of the estimation of this Figure were explained in IIIc, and some useful definitions and further details of estimations can be found in Annex 2.

[15] The values used in this estimation were obtained through a preliminary estimation of Ulyssea (2018).

Figure 10. Firm informality and cost structure

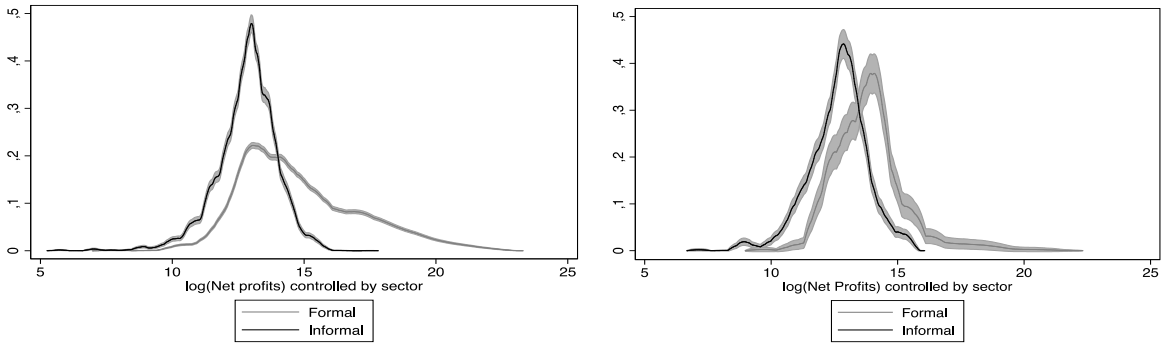


Fuente: EGG

Figure 11 estimates the net profits of firms as the value-added net of the costs described in Figure 10 and controlled by sector. The results of this estimation show that, as in the case of productivity, formal firms tend to have higher profits than informal firms, but there is a large zone of profits where both types of firms operate. This zone is even larger than in the case of the value added, which stresses the importance of a further reduction of formality cost or an increase in informality cost. However, these results are sensitive to the value of some parameters as the ones involved in the enforcement cost.

Figure 11. Net profits and firm informality

11A. Total **11B. Microbusiness**



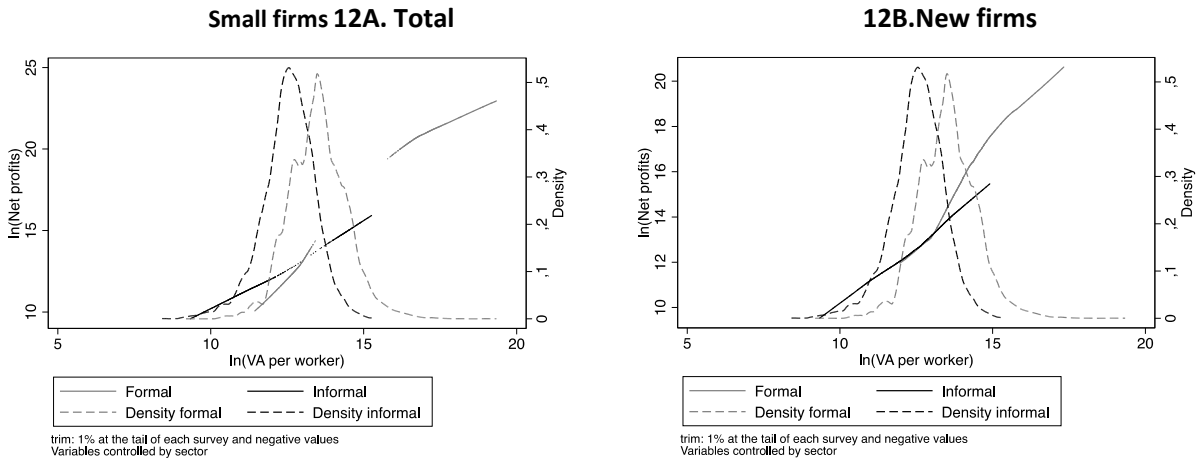
Source: EEG (2019) The estimation of new firms does not include the manufacturing industry because of lack of information.

The great coincidence between the net earnings of formal and informal firms explains the high level of informality in Colombia and suggests that there must be a certain reason that determines that some firms are formal and others informal. According to Figure 12, the heterogeneity in the productivity of formal and informal firms is related with differences in expected profits, which in turn explain the formality decision of each firm. In other words, firms with lower productivity find

it more profitable to be informal; and the firms with higher productivity find a higher profitability in being formal. The turning point occurs around COP$1.2 million per worker[16].

This interpretation is consistent with a density curve of informal firms bulkier in the segment where it is more profitable to be informal, and a density curve of formal firms bulkier where it is more profitable to be formal. This, and a certain correspondence between value added and firm size, explains why small firms tend to be informal and larger firms tend to be formal. This discussion highlights one of the basic characteristics that a theoretical model must have to understand the problem of informality: the heterogeneity of agents.

Figure 12. Net profits, productivity, and firm informality



Fuente: EEG. These figures are estimated nonparametrically with the Lowess estimator, that does not allow the use of weights. New firms are those with 1 or less year old, and do not include firms whose information comes from EAM.

The previous graphical results are confirmed in Table 3, that shows a regression between net profits, value added, informality and the interaction of the two. According to the first regression in the table (1), the relation between net profits and value added is positive, as expected; and the relation between net profits and informality is also positive, meaning that in general informal firms tend to more profitable. However, this later impact tends to diminish as the firm tends to be more productive. This equation also shows that creating a firm as a business opportunity and the education of the entrepreneur have a positive impact on profits, but for those firms, on average it is more profitable to be informal. According to equation (2), this holds also for new firms. However, when I clustered results by survey the only variable and interaction that persists significative is the value added for the whole survey, and to a lesser extend the variables related with the reason to create a firm. Figure 13 shows the relationship bet entrepreneurs' skills, the reasons to create a firm and profits.

[16] Obtained running the two variables but not controlling the productivity variable by sector

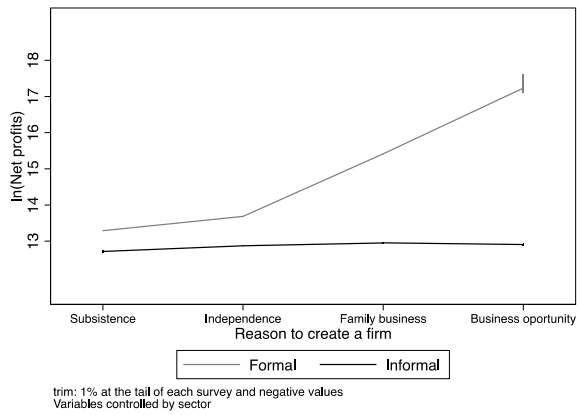
Table 3. Estimation of net profits

	(1)	(2)	(3)	(4)
	Net Profits	Net Profits	Net Profits	Net Profits
Informal firms	14,94***	10,85***	14,94**	10,85
	[15,66]	[3,99]	[5,23]	[2,82]
Value added per worker	2,125***	1,786***	2,125***	1,786**
	[31,42]	[11,02]	[10,94]	[6,65]
Informal firms # Value added per worker	-0,998***	-0,638***	-0,998**	-0,638
	[-14,20]	[-3,61]	[-5,28]	[-2,54]
Business Opportunity	0,742***	1,639**	0,742	1,639*
	[4,99]	[3,23]	[2,27]	[4,56]
Informal firms # Business opportunity	-0,833***	-1,824***	-0,833	-1,824*
	[-5,49]	[-3,49]	[-2,51]	[-4,68]
Skilled owner	0,427**	0,517	0,427	0,517*
	[2,81]	[1,03]	[1,44]	[3,53]
Informal firms # skilled owner	-0,501**	-0,446	-0,501	-0,446
	[-3,19]	[-0,88]	[-1,71]	[-3,17]
Constant	-16,47***	-12,80***	-16,47**	-12,8
	[-17,90]	[-5,04]	[-5,63]	[-3,06]
Observations	23368	727	23368	727
R-squared	0,7	0,666	0,7	0,666
Controlled by sector	yes	yes	yes	Yes
Errors clustered by survey	no	no	yes	Yes
Robust errors	yes	yes	yes	Yes
Survey	total	<=2yr old firms	total	<=2 yr old firms

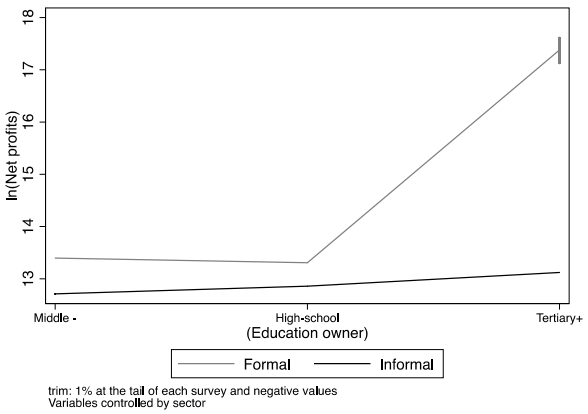
Source EEG (2019) t statistics in brackets. Weights are used. Value added and net profits are in logarithms. Data trimmed at 1% in each tale of each survey and negative VA values are not considered. Entrepreneurs of firms whose information is taken from EAC, EAM and EAS are assumed to have tertiary education (skilled owner) and being formal. New firms are those with 2 or less year old, and do not include firms whose information comes from EAM.

Graph 13. Net profits, owner skills and reasons to create a firms

13A. Reason to create a firm



13B. Owner skills



Fuente: EEG. These figures are estimated nonparametrically with the Lowess estimator, that does not allow the use of weights. New firms are those with 2 or less year old, and do not include firms whose information comes from EAM.

In sum, this section provides some evidence on the relationship between value added and profits, mediated by informality. Results are consistent with the interpretation that low productivity firms are informal because either they can't afford being formal, or because they find less profitable to be formal, whereas more productive firms find more profitable to be formal.

V. WORKERS AND WORKERS INFORMALITY FACTS

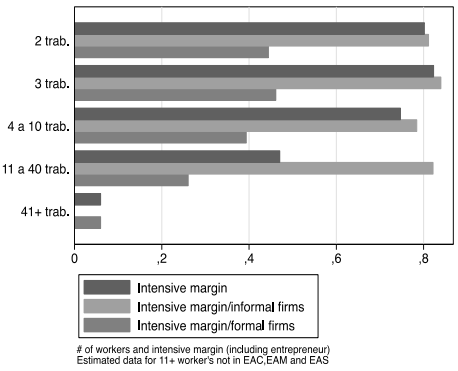
a. Intensive margin

There is an on-going discussion about the relative importance of the intensive and the extensive margin in each country. In countries as Mexico, the intensive margin receives all the attention of the literature, while in Colombia the large portion of self-employment suggest that the extensive margin can even be more important. The true is that both margins are important and even correlated, although this correspondence is not one on one, as I show in this section.

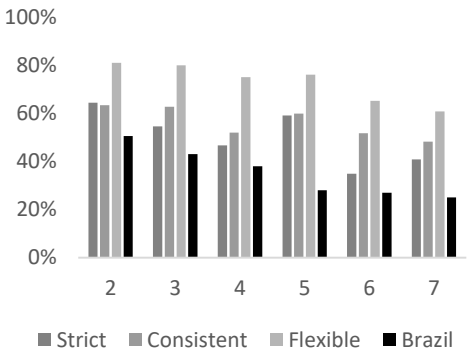
Figure 13 illustrates the extensive range of the EEG. According to this estimate, although formal firms tend to hire more formal workers (the margin is lower); informal firms also hire formal workers. This behavior is present even in the segment that does not have estimated data. In comparing to Brazil, labor informality on formal firms tends to decrease slowly and somewhat more erratically, and therefore it is not possible to assume that all workers in firms other than microbusiness are formally hired and neither that informal firms hire only formal workers.

Figure 13. Intensive margin of informality

13A. Colombia (including entrepreneur)



13B. Colombia and Brazil (excluding entrepreneur)



Fuente: EEG. Note: intensive margin of firms 40+ is not published because of number of observations.

b. Wage gap

As explained in the introduction, the wage gap between formal and informal workers and controlled by observable characteristics is often used a sign of market segmentation, because it

means that informal and formal workers are intrinsically different and perform different activities, but it can also mean that there are missing variables in describing firms or workers. One of the most interesting exercises that can be carried out from EMICRON is the determination of the income gap between formal and informal workers, controlling for firm fixed effects, and therefore limiting firm's information gaps. However, the fact that the workers skill variable in EMICRON can only be estimated reduces the scope of the exercise.

Table 4 illustrates the results of the estimation of the wage gap for the case of Colombia. Columns (1) and (2) of this table show the estimation using GEIH workers database and including and excluding the skill variable. As can be seen, the major determinants of the wage gap are labor informality, skill level, gender, age and experience and the coefficients do not change significantly with the exclusion of the skill variable. Column (3) shows the results of the same estimation but using information from EMICRON. The results are very similar to those of the GEIH, particularly regarding the coefficient of labor formality, which gives an additional value to the EMICRON estimate. As explained before, the magnitude of this coefficient could in several cases be interpreted as a market segmentation sign.

Column (4) presents the same estimate but with some firm fixed effects, such as the level of informality and the number of employees in the firm, and the employer's formality and education. These variables are significant and reduce the relevance of labor informality in determining the wage gap. Column (5) contains the same information but with all firm fixed effects. According to the results, the impact of labor informality is reduced but not eliminated, which can be explained by the lack of a skill variable. However, within formal firms (column 7) the impact of labor informality on the wage gap is eliminated when firm fixed effects are included, as is the case in Brazil (Ulyssea, 2018) and shown in Annex 3, suggesting lower support for the application of the dualist theory in Colombia. However, as I explained before results might change by including self-employment in the database.

Table 4. Wage gap between labor informal and formal workers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	GEIH	GEIH	EMICRON	EMICRON	EMICRON	EMICRON	EMICRON
Informalidad laboral	-0.3243***	-0.3682***	-0.3774***	-0.2216***	-0.1337***	-0.1693***	-0.0361
	(0.006)	(0.006)	(0.016)	(0.018)	(0.039)	(0.046)	(0.068)
Calificación	0.3352***						
	(0.006)						
Mujer	-0.1361***	-0.0979***	-0.1585***	-0.1584***	-0.0565***	-0.0589***	-0.0519
	(0.006)	(0.006)	(0.020)	(0.020)	(0.016)	(0.017)	(0.047)
Edad	0.0363***	0.0359***	0.0099***	0.0084***	0.0130***	0.0137***	0.0053
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)
Edad2	-0.0004***	-0.0004***	-0.0001***	-0.0001***	-0.0001***	-0.0001***	-0.0000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Experiencia	0.0011***	0.0012***	0.0008***	0.0009***	0.0009***	0.0008***	0.0013***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Informalidad empresarial				-0.0346*			
				(0.020)			
Informalidad laboral - jefe				0.1933***			
				(0.019)			
Educación - jefe				0.0521***			
				(0.008)			
Número de empleados				0.0214***			
				(0.002)			
Observaciones	112,874	112,874	20,319	20,319	15,509	14,064	1,445
R-2	0.375	0.375	0.353	0.374	0.951	0.953	0.863

Efectos de firma	No	No	No	No	Yes	Yes	Yes
Muestra	Todas	Todas	Todas	Todas	Existencia de brecha	Formales	informales

Fuente: GEIH (2019) y EMICRON (2019). La base de datos se restringió a aquellos que trabajan mas de 20 horas a la semana para evitar el efecto de tiempos parciales. También se replicaron los resultados incluyendo esta variable, sin obtener cambios significativos en los coeficientes.

VI. CONCLUSIONS AND RESEARCH AGENDA

Faced with a reality in which the country does not have a business census, this exercise is a first step in understanding the firm ecosystem. As shown throughout the text, it allows an adequate description of business informality and its relationship with labor informality, which can be summarized in the following stylized facts:

- Colombia's firm structure shows a high share of informal firms (93%) even after ruling out self-employment.
- Colombia has high rates of firm and labor informality in Colombia, which are decreasing in the number of workers, but at a lower rate than in the Brazilian case. This means that although there is a correspondence between size and informality levels, we cannot rule out informality among larger firms.
- Unlike the case of Brazil, where the single tax scheme provides social security protection and therefore firms under this scheme tend to be fully formal firms, firm informality in Colombia is not a binary condition, and is important to include in the analysis different informality criteria scenarios.
- There is some evidence of market segmentation at the sector level, but at the same time, formal and informal firms share similar levels of productivity. This apparent contradiction is consistent with the fact that there are several types of firms in the country: subsistence firms with few possibilities to be integrated into the formal economy because of their low productivity, firms of relatively higher productivity to whom the decision of being formal is more sensible to a cost-benefit exercise; and high productivity firms that find more profitable to be formal.
- There is also an important discussion about the source of firm heterogeneity and whether it is a cause of informality or a consequence. Ulyssea (2018) links firm's heterogeneity with a productivity unobservable variable related to entrepreneur's skills and argue that heterogeneous firms enter the market and preselect into formality or informality according to a cost-benefit analysis. This hypothesis is consistent with my finding of low-value-added per worker new firms facing more profitable to be informal and high value-added per worker new firms facing more profitable to be formal in the case of Colombia.
- However, the previous finding does not need to rule out the existence of productivity gains derived from formality documented by Perry (2007) and implemented by Alvarez and Ruane (2019) by allowing different productivity shocks to formal and informal firms.

This evidence corroborates one of the most important conclusions obtained from the empirical study of informality: the heterogeneity of agents. Indeed, informality is a complex phenomenon that covers most firms and workers in Colombia and in many Latin American countries. These agents range from the street vendor to large companies that evade taxes and workers who seek flexible work and cannot find these options in formal work. Trying to combat all types of

informality with the same recipe is not only impossible, but also inappropriate. Particularly, it is not possible to combat subsistence informality, with monitoring and control policies that can be successful in controlling tax evasion in firms of greater relative size (Loayza, 2016 and Fernández and Villar, 2017). Finding that labor productivity is key in determining the type of firm's informality makes easier the idea of implementing heterogeneous policies to heterogeneous firms.

This paper settles the basis for a research agenda on informality that should include firm's and worker's heterogeneity, and labor and firm informality. Ulyssea (2018) is a good approach to follow but it is important to release some assumptions as restricting firm and labor informality to microbusiness or forbidding informal firms to have formal workers. It also needs to be estimated for different scenarios of informality. Concerning the ex-ante productivity condition, I find it reasonable and consistent with my findings for the Colombian case, but different ex-post productivity shocks for formal and informal firms, and probably, different profit functions should be considered. Also, a big homework of the agenda is the inclusion of self-employment, but previously a detailed analysis of the data, should be performed before taking the decision of including it as a one worker firm or as a different type of firm.

Finally, as discussed in the introduction, the possibilities offered by this new database go beyond what has been discussed in this paper and having most of the observations linked to the household survey opens a new horizon of analysis in which production, informality and welfare variables are more closely related. Therefore, an analysis of the impact of small firm productivity and informality policies on welfare, poverty, income distribution and discriminated population groups, can also be a next item in the agenda.

Bilibiography

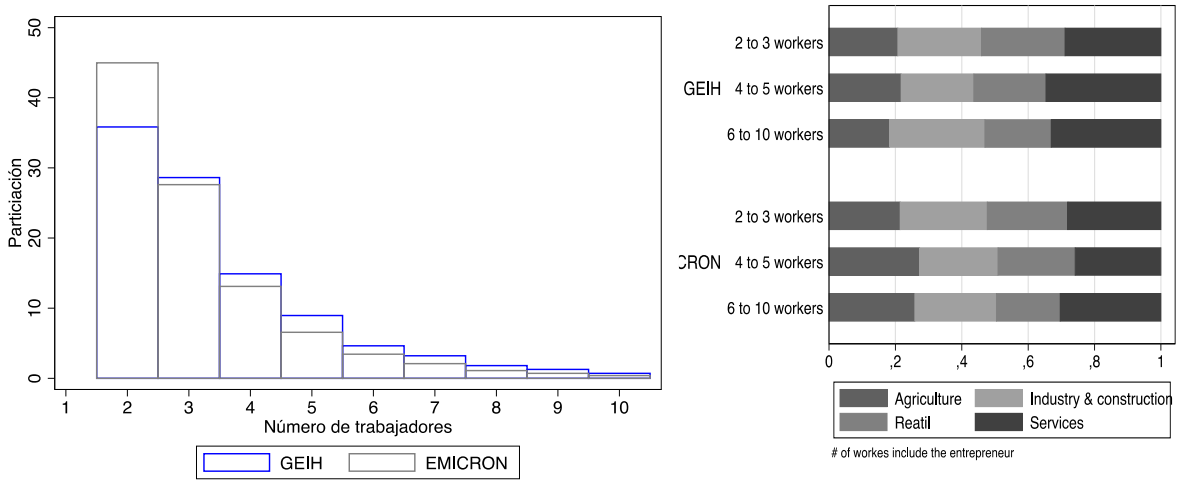
- Alvarez, J., & Ruane, C. (2019). Informality and Aggregate Productivity: The Case of Mexico.
- Cox, N. and Jenkins, S., (2011), INVGAMMAFIT: Stata module to fit a two-parameter inverse gamma distribution.
- Dane (2013) Metodología general Gran Encuesta Integrada de Hogares - EAC
- Dane (2020a) Metodología general Encuesta Anual de Comercio - EAC
- Dane (2020b) metodología general Encuesta Anual Manufacturera -EAM
- Dane (2020c) Metodología general Encuesta Micro-establecimientos - EMICRON
- Dane (2020d) Metodología general Encuesta Anual de Servicios - EAS
- Dane (2021) Censo económico de Colombia. Documento metodológico censo económico, 2021
- De Soto, H. (1989). The other path (p. 17133). New York: Harper & Row.
- De Soto, Hernando. 2000. The Mystery of Capital: Why Capitalism Triumphs in the West and Fails
- DNP (2019). CONPES 3956 Formalización Empresarial. https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3956.pdf

- Eslava M, Haltiwanger JC and Pinzón A. 2019. Job creation in Colombia versus the US: "Up or out dynamics" meets the "life cycle of plants." NBER Work. Pap. 25550
- Fernández, C. (2020). Informalidad empresarial en Colombia. Coyuntura Económica: Investigación Económica y Social. 50, 133-168, diciembre.
- Fernández, C. & Mejía, L. F. (2021). "Rigideces del mercado laboral en Colombia: tendencias, perspectivas y recomendaciones". En Fedesarrollo, *Descifrar el futuro: la economía colombiana en los próximos diez años*, capítulo 4, 261-319, Penguin Random House.
- Fernández, Villar y Gómez (2017). "Taxonomía de la informalidad en América Latina," Coyuntura Económica, Fedesarrollo, vol. 47(1 y 2), pages 137-167, December
- Fernández, Villar y Gómez (2017). "Taxonomía de la informalidad en América Latina," Coyuntura Económica, Fedesarrollo, vol. 47(1 y 2), pages 137-167, December
- Harris, J. R., & Todaro, M. P. (1970). Migration, unemployment and development: a two-sector analysis. The American economic review, 60(1), 126-142.
- Jenkins & Van Kerm, 2007. "PARETOFIT: Stata module to fit a Type 1 Pareto distribution," Statistical Software Components S456832, Boston College Department of Economics, revised 11 Nov 2015.
- Jenkins, 2007. "LOGNFIT: Stata module to fit lognormal distribution by maximum likelihood," Statistical Software Components S456824, Boston College Department of Economics, revised 01 Jun 2013.
- Levy Algazi, Santiago, 2018. "Under-Rewarded Efforts: The Elusive Quest for Prosperity in Mexico,"IDB Publications (Books), Inter-American Development Bank, number 8971
- Lewis, W. Arthur. (1954) Economic development with unlimited supplies of labour. The Manchester school, 1954, vol. 22, no 2, p. 139-191
- Loayza, N. (2016). Informality in the process of development and growth. (W. B. Group, Ed.) *Policy Research working paper* (no. WPS 7858).
- Maarten L. Buis, 2007. "HANGROOT: Stata module creating a hanging rootogram comparing an empirical distribution to the best fitting theoretical distribution," Statistical Software ComponentsS456886, Boston College Department of Economics, revised 18 May 2011.
- OECD (2007) Eurostat-OECD Manual on Business Demography, Statistics European Commission ISBN 978-92-79-04726-8, Cat. Number: KS-RA-07-010-EN-N-N, ISSN 1977-0375
- Perry G, Maloney W, Arias O, Fajnzylber P, Mason A, Saavedra-Chanduvi J. 2007. *Informality: Exit or Exclusion*. Washington, DC: World Bank
- Ulyssea G. 2019. Formal and informal firm dynamics. Unpublished manuscript, Univ. Oxford, Oxford, UK
- Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from Brazil. American Economic Review, 108(8), 2015-47.

Annex 1. Internal consistency checks

This annex compares the sectors in which the different surveys overlap to verify the robustness of the exercise, and to indicate to Dane the information gaps that exist in the system. Graph A1A shows the histogram of the EMICRON survey and GEIH (firms), and Graph A1B shows the sectoral differences between the two surveys. Unsurprisingly, the match is almost perfect, because one survey comes from the other. However, there are still some differences derived from the firm selection process in EMICRON, which corresponds to the criterion of ownership of the means of production; and whose information is not available for the entire GEIH. This selection particularly affects smaller firms.

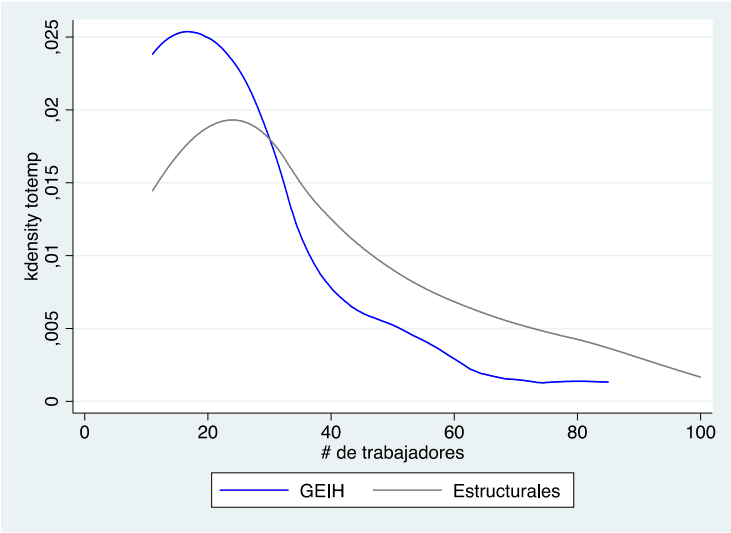
Figures A1A y A1B. Firms and informality by sector and size



Source: GEIH and EMICRON

As Figure A2 illustrates, the correspondence between the sectoral surveys and the GEIH (formal firms) is less clear, even though the GEIH sample is restricted to the relevant sectors. In particular, the GEIH seems to capture a larger number of small firms, while the structural surveys are more accurate to capture the relatively larger firms. This finding itself is of interest only to the statistical authorities and for the conduct of the Census and confirms the complementarity that exists between the two types of survey.

Figure A2. K-density curves GEIH and structural firms



Source: EAM, EAC, EAS and GEIH

Annex 2 Further definitions

- **Payroll Cost of hiring formal workers.** This cost includes social security, and vacations, severage, "cajas de compensación" and other. The social security cost of the entrepreneur is also included. I allowed informal firms to have formal workers if reported, which is not an infrequent event.

- **Sector.** The recorded sectors in this paper are 1: Agriculture and mineral extraction. 2: Industry and construction. 3, Retail; 4: Services. Note: Manufacturing includes construction in the case of Colombia.

- **Value added:** Value added is obtained directly from EMICRON and the structural surveys. Although GEIH does not provide information to estimate the Value Added with the specification suggested by Ulyssea using the GEIH, it is possible to estimate the value-added net of maintenance and functioning cost as the profits plus the wage and payroll bill and the formality cost. Formality costs were deducted.

- **Labor productivity:** Labor productivity is estimated as Value Added per worker

- **Formality cost.** Formal firms incur in costs to be formal as hiring accounting officer/lawyer/advertising licenses, association fees and chamber of commerce costs,

food licenses and carnets[17]. However, some of these costs are also pay by informal firms. To estimate the portion of this variables that is not formality cost I run a regression of these three variables on informal firms and using the coefficients to predict it on formal firms.

- ***Labor enforcement cost*** Following Ulyssea (2018) I assume the labor enforcement cost to be W $L_j^i(L_j^i)/d^i$). However, I used different parameters obtained for preliminary estimations of this model.

- **Profit taxes.** This is estimated as 0,39 (average paid tax rate by microbusiness estimated by Martinez,2018) times the likelihood that a formal firm contributes to this tax (observed in EMICRON as $\sum_1^K \mathrm{P2991}_k^f$/K times the value-added net of formal labor cost (including the entrepreneur's social security if the firm is registered as a natural person). This tax is only payable by formal firms with gross income higher than the exempted bracket in 2019 ($3'716,916), and given that $Grossincome_k^f - IC_k^f - Wages_k^f - Payroll_k^f$ >= ($Wages_k^f + Payroll_k^f$).

- **VAT taxes.** This is estimated as 0,088 (average paid tax rate by microfirms estimated by Martinez,2018) times the likelihood that a formal firm contributes to this tax in EMICRON as $\sum_1^J \frac{\mathrm{P2991}_k^f}{K}$) times the value added. This tax is only payable by formal firms with gross income higher than the exempted bracket in 2019 ($3'716,916).

- **Labor income:** Through this paper labor income includes non-monetary food and housing in estimating Wage or Pro-labor income. I In the Colombian case I also included income received in the form of home and food[18]. Labor income in the structural surveys is estimated as the total wage bill divided by the number of workers. In the EAM it is possible to estimate it for blue-collar workers.

- **Skilled workers:** I used the criteria of tertiary education to identify skill workers. Since EMICRON does not records the education level I predicted the skill variable using a logit model from GEIH (workers

Annex 3 Wage gap estimated for the case of Brazil

17 officer/lawyer/advertising; licenses, association fees and chamber of commerce costs and other cost as food licenses and carnets

18 EMICRON, labor income imputed among those workers that report social security but do not report income. The profit/income received by the entrepreneur is not included, neither this variable is measured by partners and non-remunerated workers. It also should be noted that after exclusions there still 59 remunerated workers that receive income equal to zero (income). Income received by independent workers at GEIH might refer to several months. In any case I not divided it by the number of months since I am using a monthly average, and therefore it is already divided by 12.

	log(*wage*)		
	PNAD (1)	ECINF (2)	ECINF (3)
Formal contract (dummy)	0.2864	0.2413	0.0311
	(0.007)	(0.030)	(0.080)
High skill (dummy)	0.4583	0.1373	0.0921
	(0.006)	(0.031)	(0.0519)
Male (dummy)	0.2980	0.1256	0.1793
	(0.007)	(0.035)	(0.0434)
Age (years)	0.0740	0.0674	0.0365
	(0.002)	(0.007)	(0.010)
Age squared	−0.0008	−0.0007	−0.0003
	(0.000)	(0.000)	(0.000)
Observations	60,899	4,502	2,675
R^2	0.446	0.401	0.872
Firm fixed effects	No	No	Yes

Notes: PNAD is the National Household Survey, and ECINF is the matched employer-employee data for formal and informal firms and their employees. Variable *Formal* is a dummy for formal employee; *Skilled* is a dummy for workers with at least high school degree. All regressions control for five-digit industry classification. Robust standard errors in parentheses.

Source: Ulyssea (2018)

Agradecimientos

Esta serie de documentos de trabajo es financiada por el programa "Inclusión productiva y social: programas y políticas para la promoción de una economía formal", código 60185, que conforma Colombia Científica-Alianza EFI, bajo el Contrato de Recuperación Contingente No.FP44842-220-2018.

Acknowledgments

This working paper series is funded by the Colombia Científica-Alianza EFI Research Program, with code 60185 and contract number FP44842-220-2018, funded by The World Bank through the call Scientific Ecosystems, managed by the Colombian Ministry of Science, Technology and Innovation.