Universidad del Rosario | Facultad de Economía

SERIE DOCUMENTOS DE TRABAJO

No. 320

Enero de 2025

MSMEs and Informality: A new employer - employee database for Colombia

Cristina Fernández Mejía

Universidad del Rosario, Facultad de Economía
Cl. 12c #4-59. Bogotá, D. C., Colombia
PBX: (031) 2970200 ext. 4163

facultadeconomia@urosario.edu.co
https://www.urosario.edu.co/Facultad-de-Economia/Inicio/
https://repository.urosario.edu.co/handle/10336/20291

MSMEs and Informality: A new employer–employee database for Colombia

Cristina Fernández*

Abstract

In Colombia, as in many other countries, there is a lack of comprehensive and recurring information on business demographics. This paper creates a method to generate a periodic firm-employee database that allows for the first analysis of the entire universe of firms in the country. It also creates a methodology suitable for implementation in countries with limited business information. After excluding self-employment (43% of workers) and agriculture, this paper identifies five stylized facts on business demographics in Colombia: 1. Micro-businesses account for 93% of firms, 31% of workers and 9% of the value added. This granularity can be partially explained by a tax waiver on small businesses. 2. The informality of business (from 30% to 88% under the registration or taxes criteria, respectively) is decreasing in the number of employees. This behavior can be explained because small and unproductive firms cannot afford formality, whereas larger firms find profitable to operate formally. 3. Labor informality is high (71% among firms, and 27% among workers) and also decreasing in the number of amployees. This diminishing trend, that in turn is related to small firms hiring few skilled workers, can be explained by a stronger enforcement on larger firms, plus asymmetric taxes. The after tax relative cost of hiring formally is ten times higher for small companies. 4. The incidence of poverty in micro-business is three times higher than in larger firms. 5. Overall, there are incentives for firms to remain small, but also incentives for small firms to do not use capital and skilled workers, affecting productivity and poverty.

Keywords: Informality, Firm informality, Informal labor market, Taxonomy of informality, Policy recommendations for informality, Firm size distribution, Informality and taxes, Informality and productivity.
JEL codes: J46, O17, L11, O47.

*Fedesarrollo researcher. The great support of Andrés García-Suaza in the development of this work is appreciated. María Angélica Arbeláez, Eduardo Lora, Daniel Gómez, Juan Miguel Gallego, Fernando Jaramillo, Alain Desdoigts and Luz Adriana Florez were also fundamental in developing the work. Marlon Salazar and Camilo Ríos contributed to the initial steps of this document, Mónica Ortiz provided helpful comments, and Alexander Sarango and Cecilia Suescún assisted me in the final steps. This work was funded by the United Nations Human Development Report for Colombia. I am also grateful with the National Planning Department (DNP), which financed an earlier version of this document.

1 Introduction and executive summary

The research question of this article is what are the main stylized facts of the business universe in Colombia, with emphasis on size and informality, and its relationship with productivity and poverty. The main objective is to create a methodology to periodically describe the firm-employee relationship by generating a business database in countries that do not have an economic census or comprehensive periodic surveys.

In order to achieve this objective, this paper compiles information available from the micro business survey (EMICRON) and the structural surveys carried out in the manufacturing, retail, and service sectors (EAM, EAC and EAS). Missing information was filled by using the household survey (GEIH), which provides representative data at the quarterly level on employers and their businesses, by assuming one employer per firm and one firm per employer, and populations weights of employers equivalent to firms weights. This exercise was carried out for the year 2019, considering that the year 2020 was still affected by COVID-19. The resulting database does not include the agricultural and mining sector, due to low representativeness. Government and domestic services were also excluded because they do not really belong to the corporate universe. Self-employment was left out of the discussion, not because it is not important, but because it is too important (43% of workers) and divergent. In fact, self-employment in developing countries is closer to a substitute for unemployment than a wage work (Donovan, Lu, & Schoellman, 2023), gets heavily affected by factors such as necessity, flexibility, or the existence of a business idea, and tends to show fewer transitions to salaried jobs (Fernández, 2023b). Details on the compiled database (EEG later, by its component initials: EMICRON, Structural, GEIH) can be found in Annex A.

The effort to compile this database is important for three reasons. The first is that, in accordance with the recommendations of the CONPES (Departamento Nacional de Planeación, 2019), the new database allows to understand the behavior of the entire set of Colombian firms, including the smallest and the informal ones. As a example, it was key to estimate (Fernández, 2023a) and (Fernández, 2023c). Additionally, it contributes to setting the foundations for carrying out a new economic census in the country, by estimating expected results, identifying missing information, and providing an alternative to periodic follow ups. The second is that, after performing some imputations, the new database allows for a joint analysis of the economic and social aspects of the productive sector, since most of the observations are linked to the household survey, which is the source for income distribution indicators. The third is that, the methodology used to generate this database can be applied periodically, not only to the Colombian case, but also to many developing countries with limited firm information available.

The EEG (2019) is not free of limitations, the most important of which is representativeness. Representativeness among firms with fewer than 10 workers is not an issue, since EMICRON is based on GEIH, and household surveys are an efficient way to capture small business. It also should not be an issue among the observations provided by structural surveys because they are supposed to incorporate all the business in selected sectors and size brackets. However, some consistency checks were carried out since the collection of data is not always optimal, and since the compilation of the EEG using the GEIH involved some assumptions. The EEG collects information from about 576 thousand firms, of which nearly 400 thousand are registered. This accounts for 61% of the

650 thousand firms identified as active in Colombia by the Statistical Directory of registered firms in 2019 in the relevant sectors (DEE Dane, 2023). Similarly, excluding self-employment, EEG firms employ 5.4 million workers, 62% of the total number of workers identified by the household survey in a comparable sample.

The second limitation is that, in order to make inferences in regard to labor informality, income distribution and worker vulnerability, it is necessary to make some imputations. These imputations were performed by using non-parametric techniques feed from different training bases, such as the same household database but with regard to questions asked to workers rather than employers (GEIH, 2019); the EMICRON (2019) and the 2021 household survey (GEIH, 2021, 2018 framework), which despite of providing post-pandemic information, asks employees whether the firm where they work is formal. Another variable that was missing was the value added among informal business with more than 10 workers (EEG observations obtained directly from the GEIH). This variable was calculated using the equivalence between value added and input remunerations.

The new database allows, for the first time in Colombia, to identify the main stylized facts of the entire business structure. The first of these facts is the immense prevalence of small firms with very low productivity. According to the EEG (2019), after excluding self-employment (43% of workers) and agriculture, micro firms account for 93% of the firms, 31% of the workers and 8.9% of the value added. The cause of the proliferation of small firms is still a matter to explore, but common sense suggests that it might be related to the existence of tax-exempt thresholds, the limited availability of human capital, monitoring and control that increases with firm size, high costs of using physical capital (Fernández, 2023c), an inflexible labor market, and even a history of violence which erodes confidence and makes relevant to maintain a low profile to evade crime.

Other stylized facts are a high rate of business and labor informality, that are decreasing in the number of workers but not enough to allow the assumption that all larger firms are formal. This diminishing behavior can be explained by larger firms finding more profitable to be formal (because of access to credit, access to government facilities and demand, ability to export and, visibility among others) while smaller firms not finding profitable, or not being able, to be formal. The diminishing behavior of the labor informality rate can be explained by higher enforcement in larger firms plus asymmetries in the tax scheme, that allows larger firms to obtain deductions that can not be used by smaller firms. More specifically, the after tax relative cost of hiring formally is 10 times higher in firms that do not pay taxes.

This disincentive to hire formally derives into a disincentive to hire skilled workers. On top of this, and according to Fernández (2023c) small firms face higher costs to access capital than larger firms (the average interest rate charged in 2018 was 9% for larger business, 16% for SMEs and 29% for micro-business, according to the OECD (2020)). Having few access to capital and skilled labor, there is no surprise in finding small firms to be unproductive, and their workers and heads, very likely to be poor. In fact, estimations performed using the EEG(2019) find that the incidence of poverty in micro-firms is 32% among workers and 11% among employers, barely 3 times higher that the one observed in larger firms; in accordance to what is indicated in Eslava, Meléndez, and Urdaneta (2021). Likewise, small firms show higher incidence of low-skilled, migrants and other vulnerable workers.

In the analyses carried out along this paper, there seems to be a correspondence between highly productive formal firms that hire formal and more skilled workers, and unproductive informal firms that hire informal and low skilled workers; as well as productivity levels and sectors where only formal firms operate, and productivity levels and sectors where only informal firms operate. However, these trends do not imply that Colombia is a dual economy, according to some exercises performed in this paper. This analysis is important because is consistent with firms transiting organically towards formality as they became more productive, and gives support to policies oriented to increase productivity in small firms as a way of decreasing poverty and increase formality.

The paper is structured as follows. Section 2 presents the literature review; Section 3, the main stylized facts derived from the EEG (2019); Sections 4 and 5, the relationship between firm size and business and labor informality, respectively; Section 6, an analysis of dualism in Colombia and Section 7, a reflection and some policy recommendations. These exercises are an example of the possibilities of analysis offered by the constructed database, but its potential goes beyond the scope of this paper.

2 Literature review

Perry et al. (2007), in the leading World Bank report, makes one of the first attempts to understand the relationship between business dynamics and informality at the Latin American level. According to their findings, informal businesses encompass not only small subsistence businesses, but also larger businesses that do not comply with regulations. They also suggest that considerable efficiency gains can be obtained by shifting resources from low-productivity firms to high-productivity firms. In terms of policy recommendations, they argued that some firms may benefit from lowering the costs of informality and react to an increase in the costs of being informal, but they also understood that the best policy for small firms requires actions such as access to credit, formal education, training, and business development services.

On the other hand, one of the most exhaustive exercises to describe the taxonomy of companies in the context of informality was carried out by Levy (2018) for the case of Mexico. According to the author, the main cause of low productivity is the misallocation of resources, closely related to informality and nonsalaried work. To reach this result, he carried out an exhaustive analysis at the company level, which concludes the following about informal companies; 1. They absorb a significant amount of capital and labor. 2. They can be found in all sectors of the economy and throughout the territory. 3. They are not necessarily illegal because most of them hire workers through nonsalaried labor (that is legal in Mexico), and not necessarily completely informal. 4. Most of them are very small (less than 5 workers), but not all informal companies are small. The proliferation of small businesses generates a proliferation of employers over workers, which is not always optimal. 5. It is a growing force. The composition of economic activity over time has shifted towards the informal sector. 6. They are unproductive. Firms that hire formal workers and/or are formal are more productive, but not all formal firms are more productive than all informal firms.

This analysis leads Levy (2018) to the premise that low-productivity firms absorb more capital and labor than they should, while the most productive do not receive enough resources; and this

process is enhanced by the longevity of those unproductive small businesses. Furthermore, unproductive small businesses act in socially inefficient ways. For example, they choose to be informal when they can be formal. Being informal, they remain far from the control of the authorities and, therefore, do not accept cash, do not grow, do not use technology, and do not hire formal workers. In some cases, these firms have high entry rates and low survival rates, which generate short-term jobs. This analysis was then summarized by Alvarez and Ruane (2019) to estimate Ulyssea (2018) for the case of Mexico.

More condensed and equally rich is the description of business informality carried out in Ulyssea (2018), Ulyssea (2019) and Ulyssea (2020) for the case of Brazil. The main findings of these articles are the following: 1) The intensive and extensive margin, or labor and business informality, are decreasing on the size of the companies. 2) Informal companies are on average smaller, run by people with less education, pay lower wages, and are less productive than formal companies, but despite these differences, formal and informal companies in Brazil coexist in all sectors and levels of productivity, and neither there is evidence of the missing middle, another characteristic of dualism in the economy. 3) The wage gap between formal and informal workers, characteristic of segmented economies, since it means that formal and informal workers carry out different tasks in the economy, disappears when controlling for the characteristics of the companies (fixed effects). 4) The dynamic selection process takes place in both the formal and informal sectors, but it is weaker in the latter and, therefore, ex-ante rather than ex-post heterogeneity is the determinant of the dynamics of the company.

Based on all these stylized facts, Ulyssea (2018) developed a model which integrates two views of informality that use to antagonize: the dual perspective of the market (Harris & Todaro, 1970; Lewis, 1954); and the De Soto margin, according to which informality is due to high entry costs (De Soto, 1989, 2000). Another advantage of Ulyssea (2018) is that it is a general equilibrium model that considers both the intensive and extensive margin of informality, which do not always react in the same way. However, there are some characteristics of this analysis that make it difficult to apply to the Colombian case. The first of which is the availability and richness of the administrative records, the second is the assumption that all firms larger than ten workers are formal and hire formally, which is not the case of Colombia.

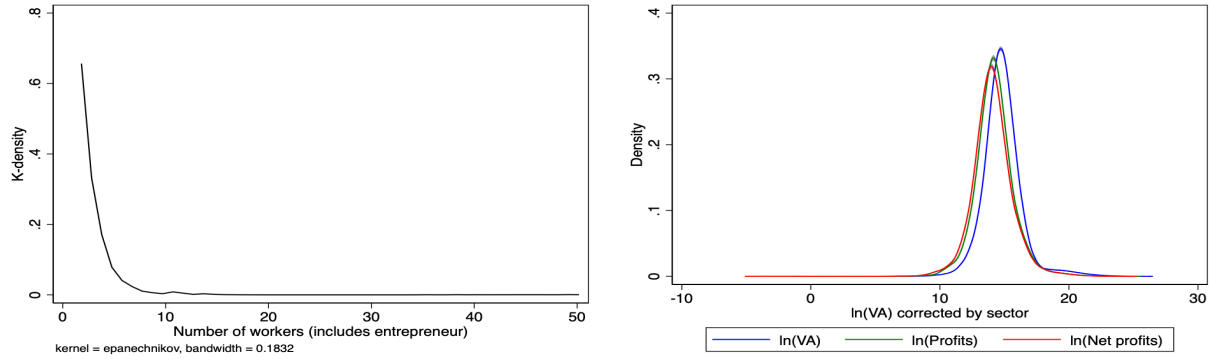
In the case of Colombia, Eslava, Haltiwanger, and Pinzón (2019) analyzed all non-micro formal manufacturing establishments and found that the size distribution exhibits a high concentration of old and small companies; pointing to a higher death rate for high-growth entrepreneurship, and a relatively high probability of long-term survival for small and likely unproductive firms. The authors consider this to be related with the low productivity growth rates observed in the country. More focused on informality, Fernández (2020) made a first attempt to understand the dynamics of business informality, but based on the Survey of Microestablishments (2013-2016), which was not representative. Although these articles made important contributions to understanding business dynamics and informality in Colombia, they were based on partial sets of information and therefore cannot address some of the main questions of this field for an aggregate level. This work seeks to fill this gap.

3 Characteristics of the Business Universe

One of the most typical characteristics of emerging economies is the granularity of business demography. According to the EEG (2019), this is also the case in Colombia: excluding self-employment (44% of total employment) and agriculture, companies with 10 or fewer workers constitute 93% of the total firms, 31% of employment (9.4 employees per company) and 9% of value added. As claimed by (OECD, 2021), Colombia shows the highest percentage of micro businesses and self-employment among member countries (the average is 80%). Figure 1a illustrates the distribution of business by number of workers. Figure 1b shows the distribution of business in according to size, value added, profits, and profits net of the cost of keeping informality away from the authorities[1]. According to this figure, the value added exhibits a normal distribution, with a long tale to the right, which is reflected in the low contribution of micro-businesses to the GDP (9%). This figure also shows how profits replicate the behavior of value added, but with less extreme data, and lower averages, indicating the impact of taxes, contributions, and enforcement on earnings.

Figure 1. Distribution of firms by number of workers, value added and profits

(a) Firms **(b) Value Added and profits**



Source: EEG (2019). Self-employment excluded.Value added and profits include confidence intervals

4 Extensive margin of informality or business informality

4.1 General characteristics of business informality

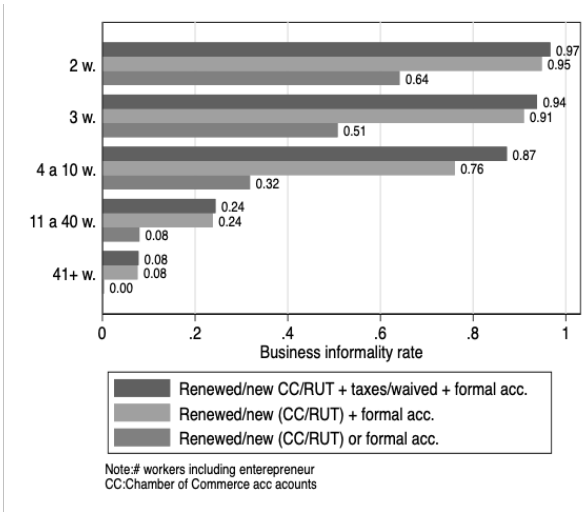
One of the problems when analyzing business informality is the existence of multiple criteria, at the point that one might account for a continuous variable, identifying compliance with a

[1]All graphs include confidence intervals and are corrected by sector differences.Profits are calculated as the net value added of salary and tax costs; and net profits, as net profits of the costs of operating informally under the control of the authorities. The median monthly earnings of firms with paid employment are close to US$1,500. To control for sector differences, the prediction errors of a regression between the logarithm of the variable of interest and the 3 economic sectors are obtained, and the regression constant is added. To avoid the impact of extreme values, 1% of the right tail of the distribution and negative values are removed from each database involved, since we work with logarithms

greater or lesser percentage of business regulations. This feature is very different in countries like Brazil where the single-tax (a scheme that considers compliance with a wide range of business regulations, including labor) prevails. Aware of these difficulties, DANE is creating a measurement methodology for business informality that considers four components: entrance, taxes, inputs, and product. Given that the methodology is still in progress, this paper uses the strictest measure of business informality (pays taxes or is exempt from paying them, has a renewed registration in the Chamber of Commerce, and maintains formal accounting) which also allows to assume that formal firms do not evade taxes. However, when the results are sensitive to the definition of informality, additional scenarios are included.

The first conclusion that can be drawn from the business informality analysis using the EEG (2019) is that it is widespread. Excluding self-employment, the strict business informality rate is 88%, and 30% of the firms do not have a chamber of commerce registration or are registered to pay taxes (RUT). As shown in Figure 2, business informality decreases monotonically with the size of the firms. However, between 8% and 24% of firms with 10 to 40 workers are informal. There are even some firms with more than 50 workers that are informal, according to the strict definition of informality (8%).

Figure 2. Extensive margin of informality (Business informality)

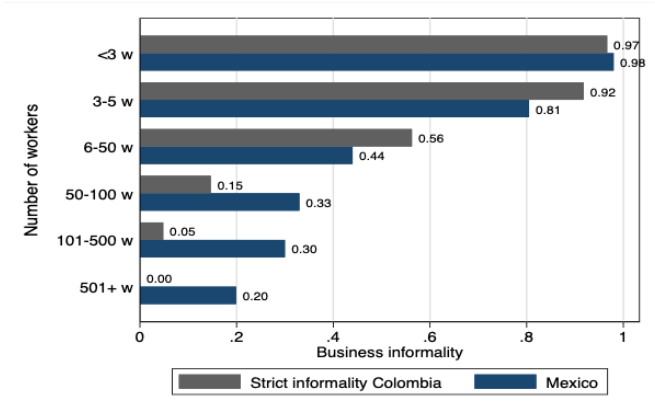


Source: EGG(2019). Self employment excluded.

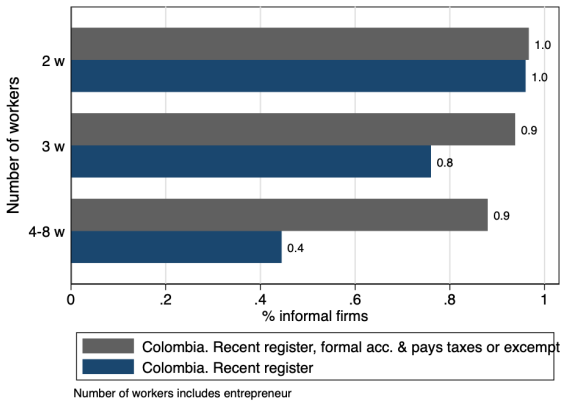
Compared to Mexico (Figure 3a), Colombia's rate is higher for businesses within 3 to 50 workers, but is lower afterwards. On the other hand, small Brazilian firms show levels of informality similar to the stricter informality criteria in Colombia, but the informality of Brazilian firms decreases faster with firm size. Therefore, assuming that firms with more than 10 workers are formal might be plausible in the case of Brazil, but not in the case of Colombia.

Figure 3. Extensive margin of informality compared to Brazil and Mexico

(a) Colombia (Strict) vs. Mexico



(b) Microbusiness Colombia (Strict) vs. Brazil



Source: EGG(2019), Alvarez and Ruane (2019) and Ulyssea (2018)

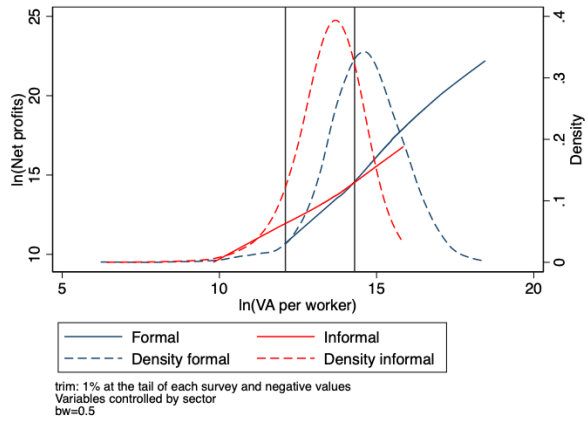
4.2 Rationality of the decreasing behavior of business informality

An explanation for the decreasing behavior of business informality on productivity is illustrated in Figure 4, that records the relationship between value added per worker, profits (solid lines) and densities (dotted lines) for formal and informal firms (blue/bold and light/red, respectively). In the first segment of the figure (less than COP$150 thousand) the profits of informal firms are positive and no formal firms operate; in the second segment (less than COP$2.2 million) the profits of informal firms are higher than those of the formal and accordingly, the informal firms outnumber the formal; and in the third segment, the opposite occurs.[2] Interestingly enough, firm distributions and profit curves cross at the same productivity level. According to Fernández (2023b), this taxonomy is not reproducible with self-employment.

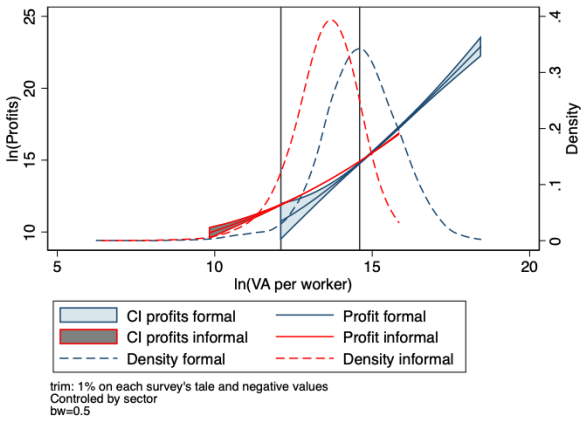
[2]Ulyssea (2018) performs a similar methodology but divides the middle segment in two, those firms with lower productivity, which he identifies as parasitic, and those that are closer to the point at which it becomes more profitable to be formal, which he identifies as DeSoto firms

Figure 4. Net earnings, value added per worker and informality

(a) Non-parametric estimation



(b) Parametric estimation



Source: EEG (2019). Excludes self-employment. Net profits exclude enforcement costs.

The previous results are formalized in Table 1, which illustrates a regression between net profits, added value, and informality as explained in Equation 1. According to the first regression (1), profits grow with the added value per worker, and on average informality tends to be a profitable alternative. However, the impact of informality on profits tends get smaller as the value added of the firm increases. Columns (2), (3) and (4) indicate that the results are robust to the inclusion of other control variables, and to the inclusion of clustered errors. It is important to note that this result, unlike Ulyssea (2018) and Fernández (2023a) that use ex ante productivity, does not imply causality, and is impossible to disguise if formality is causing a higher value added percapita or the opposite.

$$l(Netprofits_i) = c + l(VAperworker_i) + informal_i + l(VAperworker)_i * informal_i + e_i \quad (1)$$

where $Netprofits_i$ and $VAperworker_1$ are controlled by aggregate sectors (4 sectors) and $informal_1$ is a dummy variable.

Table 1. Estimate of net profits in firms with paid workers

	(1)	(2)	(3)	(4)
	Profits	Profits	Profits	Profits
Value-added per worker	1.900***	1.745***	1.745**	1.745***
	[28.22]	[25.11]	[131.06]	[11.24]
Informal firm	9.316***	6.952***	6.925*	6.952*
	[8.75]	[6.35]	[28.00]	[3.25]
Value added per worker* informal firm	-0.634***	-0.464***	-0.464*	-0.464*
	[-9.07]	[-6.46]	[-27.09]	[-2.99]
Robust errors	Yes	Yes	Yes	Yes
Controls	No	Yes	Yes	Yes
Cluster errors informality	No	No	Yes	Yes
Cluster informality and survey	No	No	No	Yes
Observations	17808	17808	17808	17808
R-squared	0.733	0.751	0.751	0.751

Source: EEG (2019). *Note:* t statistics in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Population weights are used. Value added and net benefits are in logarithms. The data is trimmed to 1% in each survey and negative VA values are not considered. Control variables: number of workers and four digit CIIU

In sum, it is possible to conclude that the informality of business in Colombia is greater and more widespread, and less of a binary condition than in the Brazilian case. It has also been shown that business informality in Colombia decreases with the size of the firms, but this decrease is not as pronounced as in the case of Brazil. For this reason, it is not possible to assume that all companies with more than 10 workers are formal. This evidence emphasizes the importance of Dane extending the EMICRON survey or developing a periodic source of information for larger informal firms. The chapter also provided an explanation for the inverse relationship between firm size and business informality,

5 Intensive margin or labor informality

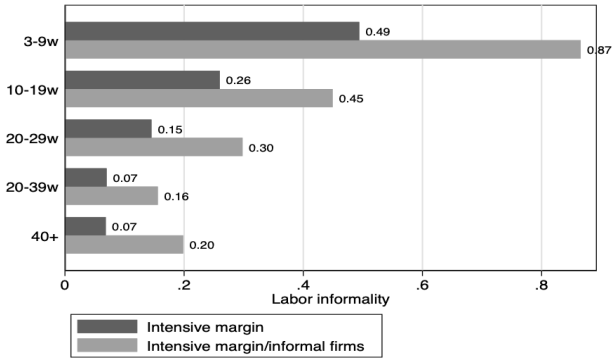
5.1 General characteristics of labor informality (intensive margin)

Although labor informality has a more straight-forward identification criteria than business informality, there is a large divergence of definitions between countries based on the internal laws (Fernández, Villar, & Gómez, 2017). The formality criterion used in this work is calculated as of health and pension contributions. The advantage of this criterion is that it allows international comparisons and does not get mixed with other topics such as the size of firms or business informality. The disadvantage is that it might identify exempted workers as informal (legal informality according to Levy (2018)).

Figure 5 shows the informality rate of formal and informal firms. The first observation that can be made from this Figure is that there are no formal firms hiring formal workers and informal firms hiring informal workers, rather the criteria of labor and business informality are intertwined. Although the average informality rate in EEG firms is 72%, the average informality rate among formal firms is 33%, and among informal firms, 84%. Formal hiring by informal firms is explained by the fact that, in order to formally hire workers, a firm does not necessarily have to be fully formal due to lack of communication among entities. Figure 5 also corroborates the decreasing behavior

of labor informality in the number of workers, but does not allows to assume that all workers in larger firms are formal, since the labor informality rate among firms with 10 to 40 workers is 19% and 6% in firms with more than 40 workers.

Figure 5. Labor and business informality

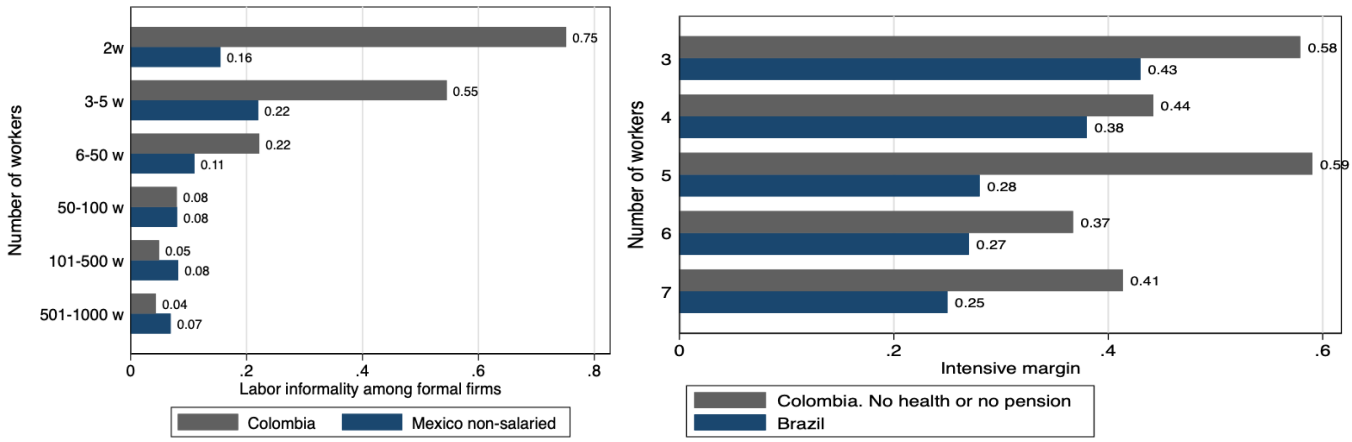


Compared to Mexico (Figure 6a), the informality of labor in Colombia is higher, but this might be related to differences in the used criteria. Compared to Brazil (Figure 6b), informality looks higher and the diminishing pattern less clear. Both statistics indicate that in Colombia it is not possible to assume that labor informality is a problem confined to micro-business.

Figure 6. Labor informality Comparisons

(a) Colombia and México

(b) Microbusinesses: Colombia and Brazil



Source: EEG (2019), GEIH (2019) and Ulyssea (2018). Thresholds:Income tax:COP$46 million, VAT:COP$100 million

5.2 Rationality of the decreasing behavior of business informality

The decreasing behavior of labor informality on size can also be explained by the rational decisions of the agents. Table 2 shows the after tax social security contribution for hiring a worker formally instead of informally. For relatively larger formal firms, it makes sense to hire workers

formally, because formal hiring costs are tax-deductible, and the amount of tax savings is only 4.7% lower to the cost of social security charged to the employer. However, if the firm has income below the exempted tax threshold (COP$46 million or US$14.000 per year) and is registered as a natural person, it cannot make any deductions; and therefore, faces a relative cost of formally hiring a worker of approximately 47% of the salary, ten times higher than in the case of larger firms.[3] This mechanism is reinforced by greater oversight and control exercised by the authorities over larger firms. A more detailed explanation of this mechanism can be founf in Fernández (2023a).

Table 2. Relative cost of hiring a worker formally and informally (2019)

Income gross annual (millions)	Income tax rate	Social security cost	Tax deduction	Difference between the cost of hiring a formal and informal worker
< 46	0%	47%	0	47%
> 46	29%	47%	$1.47 \times 29\% = 0.43\%$	$47\% - 0.43\% = -4.7\%$

Source: Own calculations, 2021 tax rates.

This tax-scheme-asymmetry impact is not marginal. According to the EEG (2019) and Table 3, 68% of the firms are below the income threshold, of which 39% are registered at the Chamber of Commerce, and of which 92% are registered under the figure of a natural person.

Table 3. Firms below the tax threshold, registered as natural persons

	Total (a)	Shares %	Registered (b)	% Registered in Chamber of Commerce (b/w)	Registered as a natural person (c)	% Registered as a natural person (c/b)
Total	576,316	100%	250,682	43%	217,639	87%
Below the threshold	391,075	68%	154,105	39%	141,075	92%
Over the threshold	185,240	32%	96,577	52%	76,564	79%

Source: EEG (2019) and own calculations. Excludes self-employment.

In sum the intensive margin is easier to analyze because of the availability of data and less diversity of criteria. The downtrend behavior observed in the data can be largely explained by enforcement and tax deductions that are only available to firms large enough to pay taxes. The high cost that face small firms in hiring formally translates into difficulties hiring skilled workers, affecting their productivity.

6 Dualism

From the analysis that has been carried out in the previous chapters, it can be deduced that formal firms are larger and more productive, and formal workers are more qualified and earn higher

[3]For 2023 the main conclusions remaining valid, but the after tax relative cost of hiring formally is being reduced even more because the tax rate has increase to 35%. In this scenario, the obvious question is why larger firms to do not hire all their workers formally. Most likely, this is related with firing cost and a lack of flexibility in formal contracts. Another recent change is the introduction of a single tax scheme (SIMPLES), whose contributions are estimated based on gross income net of social security, a middle step in terms incentives for hiring formally between the firms that do not pay taxes and the larger firms.

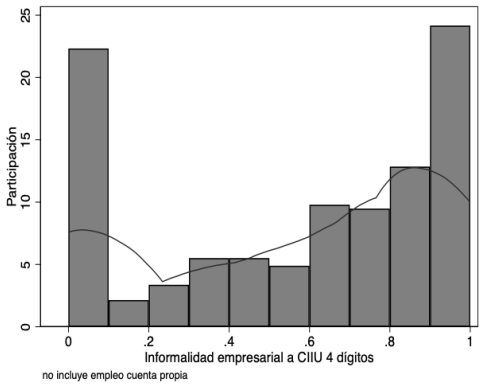
incomes than their informal peers. However, this does not necessarily imply that there are two inherently different economies with few communicating vessels (segmentation or dualism).

Identifying the degree of dualism in the economy is important because the policy recommendations diverge according to this indicator. For example, according to the theory of the dual economy, the formalization of firms and workers would by itself improve the productivity, income, and welfare of workers, which is not necessarily valid if the reasons behind the formality decision continue to be valid. On the other hand a less segmented market allows workers and firms to transit organically to formality as they become more productive. In other words, the causality between informality and productivity is at the center of this discussion. The literature has designed several tools to determine the degree of segmentation of the economy. Perhaps the most widely used is the identification of transitions between formality and informality, widely used by Maloney (2004) to establish the little segmentation of the Brazilian and Mexican markets. Unfortunately, it is not possible to replicate the exercise because the household survey in Colombia does not contain panel information.

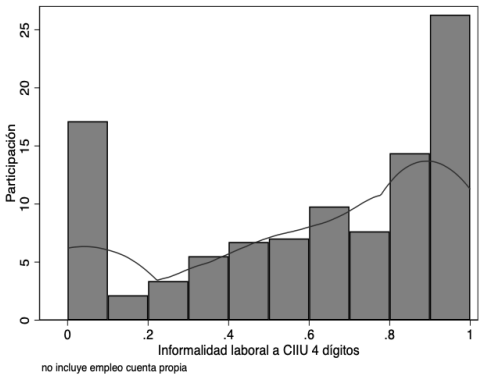
Another recurring exercise is the identification of sectors and geographical areas that are predominantly formal or informal, used by Ulyssea (2018) in the case of Brazil. Figure 7 replicates this exercise for the Colombian case, showing more segmentation than in the case of Brazil, but not complete segmentation. According to Figure 7a, the proportion of predominantly formal sectors (informality rate less than 10%) is only 22%, and the proportion of predominantly informal sectors (informality rate higher than 90%) is 23%; the remaining 55% are evenly distributed across middle rates of informality[4]. In the case of micro-business in Brazil the sectors not completely formal or informal is close to 70%. In the case of informality of the labor (Figure 7b), the segmentation is slightly less: 16% is predominantly formal and 26% predominantly informal.

Figure 7. Informality and economic subsectors

(a) Business informality



(b) Labor informality



Source: EEG (2019). Own calculations. Panels (a) and (b) exclude self-employment.

[4] A flexible criterion of informality is used and a dis-aggregation to four-digit ISIC (376 sectors). When the stricter criterion is used, naturally, the percentage of informal sectors increases to 60%, but the percentage of formal sectors is lower at 10%.

Likewise, as has been repeatedly illustrated in this work, formal firms are on average more productive and have higher profits than informal ones; and there are segments where only formal or informal firms operate, but the distributions are far from being considered dichotomous because there is a wide range of productivity where the two types of firm coexist (see, for example, Figure 4).

Finally, an exercise used by Perry et al. (2007) and by Ulyssea (2018) seeks to identify whether the differences in worker productivity (using wages as a proxy) are due to the informality of their workers. In implementing this exercise, it is possible to identify not only whether productivity differences are associated with labor productivity, but also with business informality. The results are illustrated in 2 and Table 4. Column (1) of this table shows the exercise carried out with the household survey, where it can be controlled by characteristics of the workers, all some but not all firm's fixed effects. According to the table the informality of labor has an impact of 22% on wages differences. Column (2) estimates the equation without controlling for the worker's education. This omission adds 2 percentage points to the labor informality coefficient. Column (3) makes the same estimates with the EMICRON and obtains a coefficient 4 points higher. Columns (4), (5), and (6) include some firm fixed effects such as the informality of the firm (responsible for 6% of the total effect), the labor informality of the boss and his educational level, causing a significant reduction in the coefficient of labor informality (from 0.28 to 0.22). When full fixed effects are implemented (column 7), labor informality is only responsible for 14% of differences in productivity, not controlling by skill. This indicate that there is not complete dualism in the market but a little more segmentation than in the Brazilian case Ulyssea (2018) where the coefficient of labor informality is not-significant.

$$l(wageinformal_{i_j} - wageformal_{i_j}) = workertraits_{i_j} + firmtraits_j + e_{i_j} \tag{2}$$

Table 4. Determinants of salary differences, as an indicator of dualism in the labor market

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	GEIH	GEIH	EMICRON	EMICRON	EMICRON	EMICRON	EMICRON
Labor informality	-0.2228***	-0.2339***	-0.2800***	-0.2682***	-0.2621***	-0.2165***	-0.1386***
Qualification	0.1427***						
Women	-0.1434***	-0.1362***	-0.1416***	-0.1428***	-0.1428***	-0.1466***	-0.0552***
Age	0.0261***	0.0261***	0.0274***	0.0272***	0.0276***	0.0266***	0.0119***
Age^2	-0.0003***	-0.0003***	-0.0003***	-0.0003***	-0.0003***	-0.0003***	-0.0001**
Experience	0.0106***	0.0108***	0.0096***	0.0095***	0.0100***	0.0098***	0.0127***
Bussines informality				-0.0867***	-0.0804***	-0.0564***	
skill owner +					0.0838***	0.0711***	
Boss labor formality						-0.1408***	
2to3 workers	-0.0852***	-0.0860***	-0.2078***	-0.2036***	-0.1994***	-0.1902***	
6to10 workers	0.0838***	0.0529***	0.0188	0.0132	0.0135	0.0061	
Observations	34,431	34,431	17,044	17,044	17,044	17,044	17,044
R^2	0.262	0.257	0.311	0.312	0.315	0.321	0.939
R^2	No	No	No	No	No	No	Yes
Firm fixed effects	all	all	all	all	all	all	all

Source: GEIH (2019) and EMICRON (2019).

The performed exercises make it possible to establish that the Colombian economy is more segmented than the Brazilian but it is not a dual economy. However, there is some degree of seg-

mentation and for some firms and subsistence workers it is very difficult to transit into the formal market. For an analysis of segmentation between self-employed and other firms, see Fernández (2023b).

7 Final thoughts and policy recommendations

Before generating the recommendations to reduce informality, it is worth asking ourselves the inner reason of this task, because informality cannot be an objective in itself. The most important reasons to reduce business informality are fiscal, control of unfair competition, compliance with other standards including labor and productivity. Reducing informality for fiscal reasons makes little sense because the costs of monitoring small firms may outweigh the benefits in terms of revenue, and on top of that there is a welfare cost. If the reason is control of unfair competition, it must be considered that although there are some benefits of operating in a small scale, the current regulation has also created inequities that tend to favor the largest and most productive firms. With respect to compliance with other standards, it is important to review the incentives that exist to comply with the different regulations. In the labor case, for example, it is important to consider that an increase in business formality does not necessarily imply an increase in labor formality if the firms to be formalized are small; because for these firms, it is very expensive to formally hire a worker.

Finally, reducing informality due to productivity requires rethinking. First, it must be considered that increases in productivity of small firms do not necessarily imply increases in the overall productivity of the country, because of their reduced participation in the added value; and the limited effectiveness of policies. However, it is possible to remove the constraints in accessing capital and skilled workers, amid other policies to increase productivity among small firms, and in turn, their feasibility of being formal and to provide means to the less favoured.

Some regulatory policy recommendations that imply structural changes are derived from this paper. As indicated by Levy and Maldonado (2021) the institutional apparatus should pass most of the benefit burden, which does not include aspects such as unemployment insurance or professional risks, to general taxes. In implementing this change, is important to consider that the government already finances some of the social security contributions through tax deductions. A first step towards this transition would be the establishment of a universal basic pension for low-income workers who cannot retire[5], for which it is not necessary to change the private insurance regime. For workers who earn more than one minimum wage, a progressive contribution rate could be established (Lora, Mejía, Benítez, Delgado, & Gutiérrez, 2021) or at least a flat rate. In the same line, it should be considered the obligation of register any firm with mercantile activity as a firm and not a as a natural person, which is equivalent to the eliminate the profit tax waivers in small firms. This can eliminate the distortion that incentives informal hiring, but probably should be accompanied by another relief to small firms.

Some more operational recommendations include the generation of formal cost-deduction vouchers that firms too small to pay taxes can deduct when they grow. The advantage of this measure is

[5]The financing of this pension would not require the availability of RAIS income.

that it encourages the growth of companies, which, as previously observed, is a particular problem in the Colombian case. There are also some recommendations regarding the single-tax scheme. The literature has associated this scheme with the reduction of informality in cases such as Brazil or Uruguay. However, it is important to consider the large fiscal costs and possible incentives for corporate dwarfism. Additionally, it should be kept in mind that little is achieved with the single tax in terms of well-being, if the conditions to comply with other regulations such as labor are not met. In this sense, the inclusion of social security in the single-tax, and/or the establishment of the single-tax rate on income net of salary and social security costs, can be effective solutions. Likewise, it is necessary to design additional mechanisms that encourage the growth of single-tax companies, such as the graduation of companies or decreasing tax benefits.

Another type of policy worth considering in this context is to increase productivity for equity purposes and that according to this paper can be a formality policy itself. Until recently, there was a perception that small companies did not grow and survived for long years with low levels of productivity. A recent study by Angulo (2023) indicates not only that there are some small companies that are growing; rather, it is possible to identify some common factors of these companies such as access to credit, technology, and formality (regardless of the mutual causality between this variable and productivity). However, more studies and databases are needed to be able to design what could be called a social productivity policy.

References

Alvarez, J., & Ruane, M. C. (2019). *Informality and Aggregate Productivity: The Case of Mexico*. International Monetary Fund.

Angulo, R. (2023). *Índice Multidimensional de Robustez de Micronegocios (IMICRO). Diseño y Resultados para una Estrategia de Focalización de la Secretaría Distrital de Desarrollo Económico de Bogotá.* (Tech. Rep.). Bogotá, Colombia: Inclusión SAS.

DANE. (2013). *General Methodology: Large Integrated Household Survey – GEIH.*

DANE. (2020a). *General Methodology: Annual Services Survey – EAS.*

DANE. (2020b). *General Methodology: Annual Trade Survey – EAC.*

DANE. (2020c). *General Methodology: Survey Micro-establishments – EMICRON.*

DANE. (2021). *Economic Census of Colombia: Economic Census Methodological Document.*

Departamento Nacional de Planeación. (2019). *CONPES 3956 - Formalización Empresarial.* Retrieved from https://colaboracion.dnp.gov.co/CDT/Conpes/Econmicos/3956.pdf

De Soto, H. (1989). *The Other Path: The Economic Answer to Terrorism*. New York, NY: Harper & Row.

De Soto, H. (2000). *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. Basic Books.

Donovan, K., Lu, W. J., & Schoellman, T. (2023). Labor Market Dynamics and Development. *The Quarterly Journal of Economics*.

Eslava, M., Haltiwanger, J. C., & Pinzón, A. (2019). *Job Creation in Colombia vs the U.S.:*

"Up or Out Dynamics" Meets "The Life Cycle of Plants" (Working Paper No. 25550). Cambridge,MA: National Bureau of Economic Research.

Eslava, M., Meléndez, M., & Urdaneta, N. (2021). *Market Concentration, Market Fragmentation, and Inequality in Latin America* (Working Paper No. 11). NY, NY: United Nations Development Programme (UNDP).

Fernández, C. (2022). *Firms, Informality & Institutions. The Case of Colombia* (Documentos de Trabajo No. 020598). Bogotá, Colombia: Universidad del Rosario.

Fernández, C., Villar, L., & Gómez, N. (2017). Taxonomía de la Informalidad en América Latina. *Coyuntura Económica*, *47*(1 y 2), 137-167.

Fernández, C. (2020). Informalidad Empresarial en Colombia. *Coyuntura Económica*, *50*(1-2), 133-168.

Fernández, C. (2023a). The Impact of Tax Asymmetries on Labor and Business informality. *Forthcomming*.

Fernández, C. (2023b). *Informality, Productivity & Inequality: An Analysis from the Point of View of Firms and Workers.* (PNUD)

Fernández, C. (2023c). Microbusiness and the Cost of Using Capital. *Forthcomming*.

Harris, J. R., & Todaro, M. P. (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *American Economic Review*, *60*(1), 126.

Levy, S. (2018). *Under-Rewarded Efforts: The Elusive Quest for Prosperity in Mexico*. Inter-American Development Bank.

Levy, S., & Maldonado, D. (2021). *Misión del Empleo. Reporte Resumen.*

Lewis, W. A. (1954). Economic Development with Unlimited Supplies of Labour. *The Manchester School*, *22*(2), 139-191.

Lora, E., Mejía, L. F., Benítez, M., Delgado, M. E., & Gutiérrez, D. (2021, April). *Reformas para una Colombia Post-COVID-19. Hacia un Nuevo Contrato Social* (Informes de Investigación No. 019238). Bogotá, Colombia: Fedesarrollo.

Maloney, W. F. (2004). Informality Revisited. *World Development*, *32*(7), 1159–1178.

OECD. (2007). *Eurostat-OECD Manual on Business Demography Statistics*.

OECD. (2021). *The Missing Entrepreneurs 2021*.

Perry, G. E., Maloney, W. F., Arias, O. S., Fajnzylber, P., Mason, A. D., & Saavedra-Chanduvi, J. (2007). *Informality : Exit and Exclusion* (No. 6730). The World Bank Group.

Ulyssea, G. (2018). Firms, Informality, and Development: Theory and Evidence from Brazil. *American Economic Review*, *108*(8), 2015–2047.

Ulyssea, G. (2019). *Formal and Informal Business Dynamics.*

Ulyssea, G. (2020). Informality: Causes and Consequences for Development. *Annual Review of Economics*, *12*, 525–546.

A Methodology to build the database

This appendix presents five methodological aspects that were considered for the creation of the database and its respective analysis. The first subsection presents the basic definitions used in the rest of the document, the second contains the guidelines that were adopted to build the database and standardize variables, the third explains the methodology to complete the database with imputations, the fourth presents the estimates of the productivity used in the segments of the new database that did not have this information, and the fifth and last subsection presents the methodology to measure the earnings of workers, when the information was not available.

A.1 Definitions

Before explaining the methodology used in this paper, it is worth establishing some definitions of variables that will be used from now on.

- *Firm:* Economic unit that develops a productive activity of goods or services, in order to obtain an income, acting as owner or lessee of the means of production[6]. Only firms with more than one worker are included. Firms with unpaid workers are usually excluded[7].
- *Firm size:* Number of workers. Includes direct workers, direct contractors, unpaid workers, partners, and employers.
- *Formal firm:* The most used definition in this work is strict: recently registered firms, with formal accounting, that pay taxes or are exempt to do it. Other definitions used include: i) registered (recently or not) and with formal accounting, ii) registered (recently or not) and iii) registered (recently or not) or with formal accounting.
- *Formal worker:* contributes to health and pensions.
- *Skilled worker:* worker with secondary / higher education.

A.2 Construction of the business universe

The EEG (2019) business universe is constructed as follows: companies with fewer than 10 workers are represented by EMICRON; companies with more than 10 workers (11+) are represented by structural surveys (EAM, EAS, and EAC)[8], and by the GEIH if they are informal or belong to a sector or size not covered by structural surveys. Indeed, under the assumption that each employer represents a firm, and given that the employer sample is representative of the total number of employers in the country on a quarterly basis, it is possible to assume that the sample of

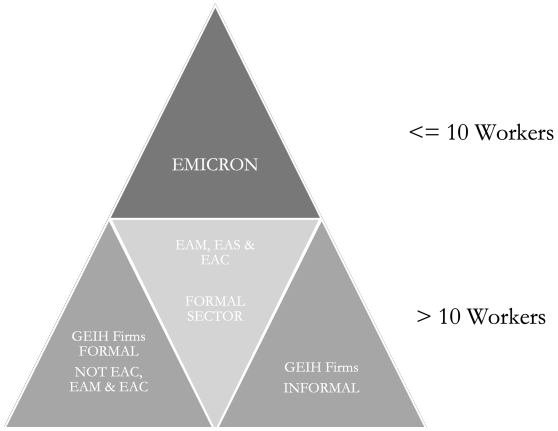
[6] DANE's definition of microbusinesses – EMICRON (2019).

[7] According to DANE and the OIT, self-employed units are those that do not have paid workers, and therefore self-employment is not the same as one-worker unit. Whereas the EEG including self-employment better describes the distribution of firms, the EEG excluding self-employment better describes the behavior of firms, because self-employment often shows choices more related to occupational choice

[8] Companies with 10 or fewer workers are discarded in these surveys.

employers is representative of the total number of businesses in Colombia[9]. It is also assumed that the companies referenced in the structural surveys are formal, since they filled out their registration number. Figure 8 shows the sources of EEG information, and each of the individual sources is detailed below.

Figure 8. EEG Information Sources (2019)



Source: author's own estimates.

- **EMICRON (Microbusiness Survey).** This cross-sectional employee-employer survey is a module of the Household Survey and, therefore, representative at the level of microentrepreneurs. Dane adjusted population weights to the universe of micro-entrepreneurs and to the loss of information in the data collection process. Its coverage excludes government business, air transport, the financial and insurance sector, and some geographical areas (DANE, 2020a)[10]

- **Household Survey (GEIH, Dec 2018 – Nov 2019).** This cross-sectional survey is representative on a monthly basis for large aggregates and on a quarterly basis, for more disaggregated levels. The part of the household survey that is most actively used in this exercise is the set of questions asked about self-employed workers (DANE, 2013)[11]. To make this survey compatible with EMICRON, the months of analysis were adjusted[12], and the sectors not covered by EMICRON were excluded.

- **Annual Manufacturing Survey (EAM, 2019).** This is a structural and exhaustive survey

[9]The jobs of the second employers are not considered in the exercise, because the variables that describe the job are different and are generally not available.

[10]EMICRON (and therefore the GEIH information used) includes microenterprises in 24 cities and their respective metropolitan areas, but does not include other territories on the periphery.

[11]Initially, the GEIH has 337,507 observations, which represents 22,303,304 workers (average for the whole year, pesos/12). However, it is necessary to exclude some sectors to make it comparable to EMICRON: The following sectors were excluded: government (12743), public services (3447), air transport (230), financial and insurance sectors (4815), and household activities that act as employers (12002).

[12]EMICRON is carried out one month after the household survey, and therefore the GEIH used corresponds to December 2018 and January – November 2019

that considers all manufacturing establishments with 10 or more than 10 people employed nationwide

- **Annual Trade Survey (EAC, 2019).** This is a comprehensive structural survey that considers all retail economic units with 10 or more people employed nationwide, exempting used-goods retailers (auto resales included) (DANE, 2020b).

- **Annual Services Survey (EAS, 2019);** It is a comprehensive and structural survey that considers all formal economic units of service at the national level. Unlike EAC and EAM, the survey coverage cutoff level differs with the subsectors[13].

- **Household Survey (2021), 2018 framework** This database is not used directly, but to estimate variables such as the intensive margin of formality, or the probability that a formal firm or an informal hires vulnerable workers. In fact, despite the fact that this information was collected after the pandemic, it has the advantage of asking each employee if the firm where they work is formal or informal

The observation unit used in this database is the "economic unit", in line with what is suggested by OECD (2007). Company size in terms of workers includes direct workers, direct and indirect contractors, interns, unpaid workers, and partners and employers[14]. In addition to sectors not included in EMICRON, such as finance, government, and household services, this analysis excludes the agricultural sector, for the sake of sectoral representativeness. In fact, the number of workers in the agricultural sector in the EEG (2019) was only 30% of what is reported in the GEIH in firms with more than 10 workers.

The database was generated for 2019, which is the first year for which there is complete information on the required surveys. This information is also available for 2020, but it has limitations in the collection of information and is altered by the effect of COVID-19. In the event of a new Economic Census, the population weights used may be adjusted to ensure the representativeness of the survey. All the household surveys' information uses the new expansion factors provided by DANE with the new survey framework that has population projections based on the last 2018 census.

Table 5 summarizes the data used in this document: 36 thousand observations were used, of which 16 thousand come from EMICRON, 19 thousand from the structural surveys (EAM, EAC, and EAS) and 458 from the questions asked of independent workers in the GEIH. These observations represent 1 million companies and close to 6 million workers. Excluding self-employment, 576,316 firms and 5,402,456 are represented in EEG (2019)[15]. Once the business universe was

[13] A detailed list of cut levels by sectors can be found in DANE (2020c), https://www.dane.gov.co/index.php/estadisticas-por-tema/servicios/encuesta-annual-de-servicios-eas

[14] This unit generally refers to the establishment, which is consistent with EMICRON, EAS and EAC. However, the information on the number of workers in the household survey and in the EAM refers to the unit that includes branches, etc.

[15] It is important to note that whereas the information collected from the EMICRON and GEIH use population weights, the information in the structural surveys is collected using a census approach, and therefore the population weights are equivalent to 1.

built, some variables were adjusted to standardized definitions[16].

Table 5. Composition of the EEG database (2019)

		EMICRON >1 worker	EAC, EAS, EAM >10 workers	GEIH (independent workers survey, firms with >10 workers)	Total
Including self-employment	#Observations	16,082	19,451	458	36,220
	# of firms	963,836	19,451	24,307	1,007,594
	# of workers	2,605,617	3,168,968	603,450	6,378,036
Excluding self-employment	#Observations	9,264	19,451	441	29,156
	# of firms	533,602	19,451	23,262	576,316
	# of workers	1,650,729	3,168,968	582,758	5,402,456

Source: EEG (2019). Population weights framework 2018. Self-employment refers to firms with unpaid workers.

A.3 Representativeness analysis

A first approximation to estimate the representativeness of the EEG (2019) is to divide the analysis in each of the subcomponents of this database. The EMICRON subcomponent does not have a business universe record, because surveys and censuses at this level are costly and ineffective. However, this survey is estimated as a module of the household survey, which is considered a good practice for counting small businesses. Likewise, the census nature of the structural surveys (EAC, EAM, and EAS) is supposed to clear all existing doubts about their representativeness. Finally, in terms of the representativeness of firms with more than 10 workers that are not represented in structural surveys, some doubts can arise from using the weights of independent workers as proxies of the firm's weights. Since these weights are designed with a criterion related to availability of information, this should be the case, but nevertheless a representativeness analysis was performed based on both the number of workers and the number of firms.

To revise the representativeness of the in terms of the number of workers, the following exercise multiplies the number of firms and the number of workers in each firm, according to EEG (2019). Similarly, the total number of workers in the GEIH (22.6 million) is restricted to the universe of the EEG by excluding firms with fewer than 10 workers (15 million) and those with more than 10 workers performing government or agricultural activities (2 million). As shown in Table 6, there are 3.7 million workers reported in the EEG (2019) compared to 5.3 million estimated by the GEIH, a ratio of 71%[17]. By sector, this ratio is 82% in services, 72% in retail, 53% in industry and 37% in construction. By firm size, a good representativeness of the number of workers is observed in firms with more than 50 workers, but in firms between 11 and 50 workers, the representativeness is low (40%). This exercise does not differentiate failures in the representativeness of the GEIH firms from the structural ones, because the GEIH (2019) does not indicate if workers are hired in formal or informal firms. Therefore, it is not possible to identify recollection problems in the collection of the structural surveys or failures in the answers of independent workers among mid-size firms.

[16]The homologation charts are available upon request.

[17]When comparing the whole survey, excluding self-employment, the ratio is 61%, which means that there are still some differences between the criteria used by EMICRON and the selected sectors in the GEIH.

Table 6. Representativeness of the EEG in firms with more than 10 workers (thousands)

	EEG	GEIH workers	EEG/GEIH
Services	2378	2901	82%
Trade	637	880	72%
Manufacturing	578	1086	53%
Construction	178	478	37%
from 11 to 50	665	1,653	40%
from 51 to 100	326	466	70%
101+	2781	3226	86%
Total	3,772	5,346	71%

Source: EEG (2019) and GEIH (2019).

Comparison of EEG (2019) with other sources of business information at the firm level is difficult due to different delimitations of coverage in terms of sectors and due to different definitions of the firm. The Directory of Companies (DEE) collected by DANE restricts its coverage to companies with at least one formal employee who performed an economic activity during the year and, in this framework, finds a total of 740 thousand active companies in 2019 (DANE, 2021)[18]. The results, which are presented in Table 7, indicate a coverage of 61%. In fact, the EEG counts 396 thousand registered units and the DEE 649 thousand units. This similarity is achieved despite the fact that the data collection procedures and sources are totally different.

Table 7. Comparison of the number of firms with other sources of information.

	EEG	EEG Registered	DEST	Registered EEG/DEST
Retail	169	126	230	55%
Services	218	154	269	57%
Manufacturing	104	75	82	91%
Construction	83	41	67	61%
Total	574	396	649	61%

Source: DEST (2019), GEIH (2019) and EEG (2019, excluding self-employment).

A.4 Imputation of missing variables

Unlike other exercises that estimate the number of companies, such as the DEE or administrative records that only account for a few characteristics of the firm, the EEG has a good set of variables[19]. Table 8 presents the observed (O), assumed (A), imputed (I), and unavailable (ND) information in the EEG (2019) and in the databases used to estimate the imputed information: 1) the GEIH household survey (2019) in its segment aimed at workers, for firms with more than 4

[18]The pre-count of the census that refers to the economic units is 2.5 million firms. However, the data is not comparable to the EEG due to differences in the definition of the firm.

[19]As in Fernández (2023b) this set of variables can be increased to include poverty, income distribution, women (%), migrants (%), youth (%) and other vulnerable populations and using the same methodology presented in this section

workers; 2) the EMICRON database (2019) at the worker level, and 3) the GEIH household survey (2021, 2018 framework). Note that the advantage of these two last databases is that they contain simultaneous information on the characteristics of the workers an employers, and the formality and informality of the firms[20].

Table 8. Availability of information to estimate productivity at the firm level

		EEG (2019)					Training databases		
		EMICRON Firms	EAM	EAC	EAS	GEIH Firms	GEIH Workers (2019)	GEIH Workers (2021)	EMICRON Workers (2019)
Worker characteristics	Labor informality	O	I	I	I	I	O	O	O
	Labor income (%)	O	I	I	I	I	O	O	O
	Skill level (%)	I	I	I	I	I	O	O	NA
	Value added per worker/wages	O	O	O	O	I	O	O	O
Firm characteristics	ISIC (2digits)	O	O	O	O	O	O	O	O
	Area	O	NA	O	O	O	O	O	O
	# workers	O	O	O	O	O	NA	NA	O
	# worker (range or s)	O	O	O	O	O	O	O	O
	# Salaried, partners & non-remunerated	O	O	O	O	O	O	O	O
	Business informality	O	O	O	O	O	NA	O	O

Source: EEG (2019), GEIH (2019), EMICRON (2019) and GEIH (2021, Framework 2018).

To carry out the imputations of the sociodemographic characteristics of the workers, artificial intelligence methods were used, and more specifically, the "random forest" method. This method estimates the probabilities that a worker has a certain characteristic from decision trees whose nodes correspond to the characteristics of the firms, such as the total number of employees, the geographic area, and the operative sector, its formality status, and if the data is post-pandemic. One advantage of the random forest is that it is possible to have data and variables omitted in some of the training bases, which is a clear difference from a traditional regression, and is key when you want to impute variables using different training bases. These estimates assume that the probabilities of occurrence of the dependent variables are the same at the level of workers and firms, conditional on certain characteristics of the firms.

Table 9 shows the results of the estimation after imputing the missing data, multiplied by the number of workers in each firm and compared with the same statistics in the household survey (DANE, 2020b), excluding the government and agriculture of one firm worker. It also shows some robustness indicators of the random forest methodology, such as the error rate (ER) and the out-of-bag error (OOB). According to the table, the imputations reproduce well the distribution of variables in the GEIH, among the relevant sectors. In fact, the only case where the null hypothesis of equal means is rejected is the average skill level in firms with more than 50 workers, which is overestimated in the EEG by approximately 1pp.

[20]The EMICRON contains information about business informality because it is an employee-employer survey, but it is restricted to firms with less than 10 workers, which is an impediment when it comes to estimating what happens in larger firms.

Table 9. GEIH and EEG (estimated for some segments) by firm size

Labor informality (training: EMICRON (2019 & GEIH (2019 & 2021)

	GEIH	EEG	Difference	Std. error	t	$P < \|t\|$
11 to 50 workers	0.221	0.197	-0.024	0.0139256	-1.73	0.084
51 + workers	0.057	0.052	-0.006	0.0034289	-1.68	0.094
OOB	0.090					
EEG	0.075					

High school diploma, training GEIH (2019 & 2021)

	GEIH	EEG	Difference	Std. error	t	$P < \|t\|$
2-3 workers	0.527	0.519	-0.008	0.005	-1.58	0.114
4-10 workers	0.635	0.643	0.009	0.008	1.10	0.269
11 to 50 workers	0.765	0.785	0.021	0.012	1.69	0.09
51 + workers	0.878	0.895	0.017	0.005	3.47	0.001
OOB	0.228					
EEG	0.2					

Labor income (observed in EAC, EAS and EAM, training: GEIH salaried (2019) & EMICRON (relevant sectors)

	GEIH (relevant sectors)	EEG (relevant sectors)	Difference	Std. error	t	$P < \|t\|$
11 to 50 workers	13.86	13.86	0.004	0.1700	0.868	0.868
51 + workers	13.98	13.88	-0.098	0.0604	-1.63	0.103
OOB	0.34					
MAE	0.46					
EEG	0.31					

Source: Own estimations using EEG (2019, population weights of workers), GEIH (2019) EMICRON (2019) GEIH (2019) and GEIH (2021). Estimation variables: size by ranges, area (when the estimate does not involve structural surveys), 2-digit sector, paid worker or partner, post-pandemic dummy, flexible firm formality (does not contribute to health and pensions) and strict (contributes to health and pensions), contributions of the head of the firm, in the estimation of business informality and education of the head in income estimation. Estimation method: random forest. ER: error rate. OOB: error outside the basket.

A.5 Value-added of observations obtained directly from the GEIH

DANE calculates the added value of the EMICRON firm and in structural surveys as gross production minus intermediate consumption (equation 3). However, this information is not requested from entrepreneurs in the GEIH. To calculate it, this article uses the added value estimate from the point of view of factor remuneration, according to which gross production is equal to disposable income, which in turn is equal to workers' remuneration, gross surplus (GS), mixed income (Π_i), and taxes on production (equation 4). Formally, the exercise can be explained as follows:

$$VA_i = GrossProduct_i - IntermediateConsumption_i + NetTaxes_i \tag{3}$$

$$VA_i = FactorRemuneration_i + TaxesOnProduction_i$$

$$VA_i = WorkersCompensation_i + GS_i + MixIncome_i + TaxesOnProduction_i \tag{4}$$

$$VA_i = (1+\tau_w)w_{if}l_{if} + w_{ii}l_{ii} + GS_{ISIC}VA_i + \Pi_i + \tau_{vat}VA_i$$

$$VA_i(1-\tau_{var}-GS_{ISIC}) = (1+\tau_w)w_{if}l_{if} + w_{ii}l_{ii} + \Pi_i$$

$$VA_i = \frac{(1+\tau_w)w_{if}l_{if} + w_{ii}l_{ii} + \Pi_i}{1-\tau_{var}-GS_{ISIC}}$$

$$VA_i = \frac{(1+\tau_w)w_iL^s(1-P^{li}) + w_iL^s(P^{li}) + \Pi_i}{1-\tau_{var}-GS_{ISIC}} \tag{5}$$

Where:

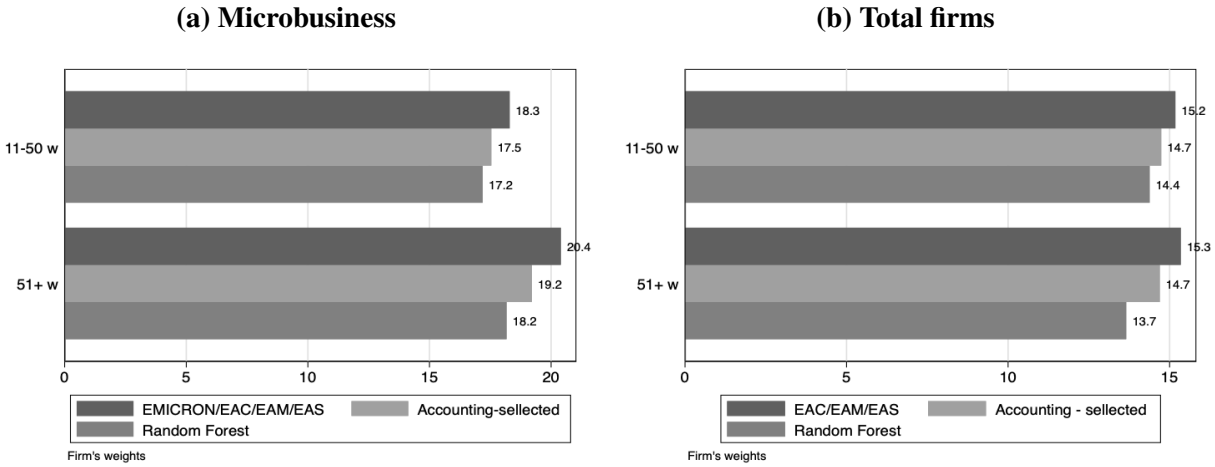
- t_{vat}: value added taxes. $t_{vat} = 0$ if $AnnualIncome < 100M$.
- GS_{ISIC}: gross operating surplus at the ISIC 4-digit level.
- L^s: total salaried workers.
- P^{li}: probability of a worker being informal (estimated).

EEG (2019) provides all the variables needed to estimate equation 5, except gross operating surplus (GS), which represents capital income that is not included in mixed income, but is estimated according to the income-use matrix (DANE). This matrix has 61 groups of activities, which can be correlated with the four-digit ISIC categories[21]. The GS is assumed to be zero in companies whose heads earn less than 100 million pesos per month.

Figures 9a and 9b show the comparison of the value added and productivity estimated in logarithms between firms with more than 10 workers. The upper bar refers to the labor productivity data observed in the structural surveys (EAC, EAM, and EAS). The middle bar refers to the calculation carried out in an accounting manner and the lower bar to the estimate of labor productivity with the artificial intelligence method described above.

[21]It is assumed that all the ISIC activities that belong to a group contribute to the gross operating surplus in an equitable manner, which is why each ISIC will be charged with the average of the group of national accounts distributed among all the ISIC that belong to each group. one of the 61 categories of National Accounts activities

Figure 9. Estimation of the logarithm of the added value with two alternative methods



Source: EEG (2019) and their own calculations based on EMICRON, EAC, EAM and EAS.

A.6 Firm profits

Firms' gross profits are equal to the sum of value added, VAT, income tax net of deduction of formal payroll (as is the case in Colombia) and formal and informal payroll, as illustrated in equation 6[22].

$$\pi^{f,i} = (1-\tau_y)\Big\{(1-\tau_{vat}VA_i - (1-t_w)w_i l_i^f\Big\} - w_i l_i^i \tag{6}$$
$$\pi^{f,i} = (1-\tau_y)\Big\{(1-\tau_{vat}VA_i - (1-t_w)w_i L^s(1-P^{li})\Big\} - w_i L^s(P^{li})$$

where

- τ_y: income tax (28%), 0% if the firm earns less than 46M per year or the firm is informal.
- τ_{vat}: VAT (19%), 0% if the firm earns less than 100M per year or the firm is informal.
- t_w: social security contributions (47%)
- w_i: average salary of the firm for salaried workers.
- l_i^f: formal workers.
- l_i^i: informal workers.
- L^s: total salaried workers.

[22]Partner earnings are assumed to be part of profit and unpaid workers receive no income.

- P^{li}: intensive margin, probability that a worker is informal in the firm (estimated in EMICRON and GEIH and assumed in structural surveys).

Equation 7 illustrates the estimation of firm profits net of the costs implied by the monitoring and control activities by the authorities. Like in Ulyssea (2018), this is a cost function, quadratic on informal labor, that reflects the increasing cost of remaining out of sight of the authority, mediated by a parameter specific to the type of business and informality of the labor[23]. Note that, unlike Ulyssea (2018), informal firms are allowed to hire formal workers. This is particularly important if the informality criterion is strict or includes the payment of taxes. In other words, the firm can hire informally if it is registered in the Chamber of Commerce, even if it does not pay taxes or is exempt from them.

$$\pi^f = VA_i - (1-\tau_w)w_i L^s(1-P^{li}) - \frac{1-L^s(P^{li}P^c)}{b_c} w_i L^s(P^{li}P^c) - \tag{7}$$
$$L^s \frac{P^{li}(1-P^c)}{b_{nc}} w_i L^s P^{li}(1-P^c)$$
$$\pi^i = (1-\tau_y)\Big\{(1-\tau_{vat})VA_i - (1-\tau_w)w_i L - s(1-P^{li})\Big\} - \frac{1-L^s(P^{li}P^c)}{b_i} w_i L^s(P^{li}P^c) -$$
$$L^s \frac{P^{li}(1-P^c)}{b_i} w_i L^s P^{li}(1-P^c)$$

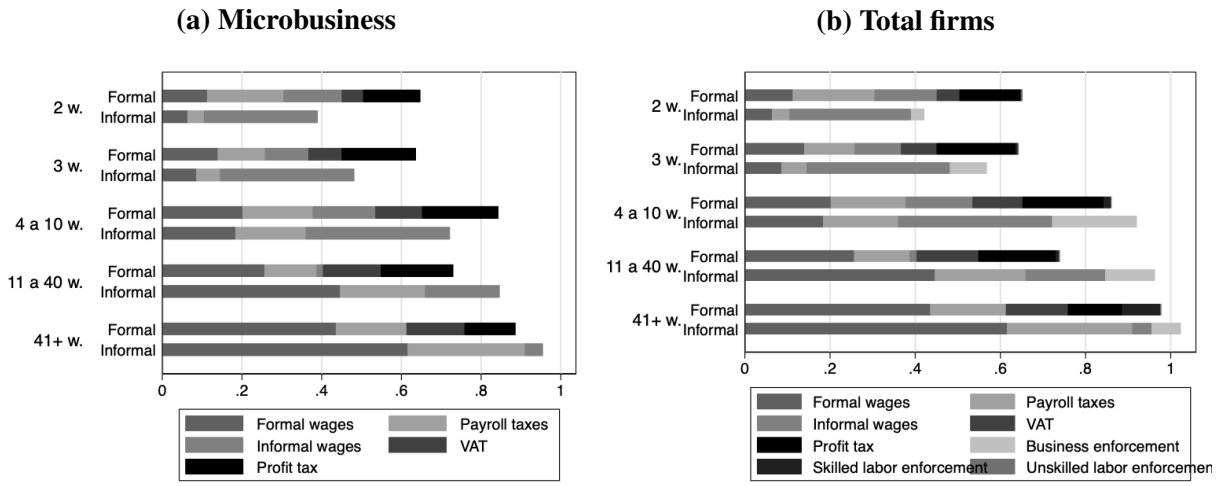
where

- π_i^f: earnings of formal firms.
- pi_i^i: earnings of informal firms.
- VA_i: value-added.
- b_c: enforcement cost parameter, skilled workers, higher = lower cost.
- b_{nc}: enforcement cost parameter, unskilled workers, higher=lower cost.
- b_i: enforcement cost parameter, informal firms.
- P^c: probability of having a skilled worker.

Figures 10a and 10b show the cost structure of firms: relative salary, regulatory and tax costs of formal and informal firms without including and including the costs that the monitoring and control of the authorities entail for the firm. As is possible to observe in the results, the costs of operating formally for smaller firms are higher, while the costs of being informal for larger firms are higher; but enforcement moves the switching point to a smaller firm size.

[23]For this calculation, the used parameters were obtained from Fernández (2022).

Figure 10. Estimation of costs other than inputs from formal and informal firms



Source: EEG (2019). Self-employment excluded