Immigration and Worker Responses Across Firms: Evidence from Administrative Records in Colombia*

Lukas Delgado-Prieto†

Job Market Paper

September, 2023

[Click here for latest version]

Abstract

Many migrants are moving to developing countries where small firms are prevalent. The interaction between firms, workers, and labor supply shocks in these contexts is mainly unknown. To answer this, I exploit the mass arrival of migrants from Venezuela in Colombia and use administrative employer-employee data covering the universe of formal workers and firms to follow natives' labor market outcomes over time. After migrants arrive, I find a reduction in individual employment concentrated at the bottom of the wage distribution (among self-employed and minimum-wage earners) and within low-paying firms. Besides, I find a negative wage effect driven by workers from the upper part of the wage distribution who work in small firms. Consistent with this, I construct a model of heterogeneous firms to show that employment and wage effects vary depending on the type of firm the worker is employed. Next, I identify the subgroups most affected by immigration by implementing a machine learning method. This method shows that firm-specific pay premiums explain the negative effect on employment and wages more than other worker characteristics. Overall, these results suggest that firms play an influential role in determining the impact of immigration on workers' outcomes.

Keywords: Immigration, Formal labor markets, Causal forest.

JEL Codes: F22, O15, O17, R23.

*I am extremely grateful to Jan Stuhler for his guidance and support in this project. I am also thankful to Michael Amior, Jaime Arellano-Bover, Agostina Brinatti, David Card, Juanjo Dolado, Jesús Fernández-Huertas, Patrick Kline, Luigi Minale, Jesse Rothstein, and Jon Piqueras for their comments and suggestions. In general, I want to thank participants at the EALE Conference, ENTER Jamboree, the 1st HUMANS LACEA Seminar, the Labor Lunch Seminar at UC Berkeley, the IRLE Seminar, the UC3M Applied Reading Group, the 23rd IZA Summer School in Labor Economics, and the Junior Seminar of the Economics of Migration for the helpful discussions. I am indebted to the Ministry of Health and Social Protection of Colombia for providing access to their data. This paper was previously circulated as: "The Role of Workers and Firms in the Impact of Immigration".

†Department of Economics, Universidad Carlos III de Madrid, Spain. Email: ludelgad@eco.uc3m.es.

1 Introduction

In the last decade, several countries across the globe have experienced large population exodus.[1] The majority of these migrants are leaving for closer destinations that, in most cases, are developing countries where small firms tend to dominate (McKenzie, 2017). Studying the labor market impacts of immigration is gaining more relevance in these contexts. However, the migration literature mostly uses cross-sectional data to identify, for instance, the wage and employment effects of immigration. This is misleading if a specific set of workers move out of employment or to other regions (Borjas and Edo, 2021; Dustmann et al., 2023). The effects are also expected to be uneven across natives as migrants can disproportionally concentrate in small firms (Delgado-Prieto, 2022).

In this paper, I study the labor market impacts of one of the most significant episodes of immigration in recent history, the Venezuelan mass migration to Colombia, and use novel administrative data that covers the universe of formal workers and firms in the country.[2] Exploiting the unequal arrival of migrants across local labor markets, I quantify the worker-level effects of immigration and construct a proxy for the role of firms in these effects.[3] A growing number of studies analyze how firms shape, for instance, wage inequality (Card et al., 2013), the gender and racial wage gap (Card et al., 2016; Gerard et al., 2021), or immigrant assimilation (Arellano-Bover and San, 2020), but less is known about how firms determine natives' adjustments to labor supply shocks.

To my knowledge, this is one of the first papers to study the impact of immigration in developing countries equipped with panel administrative data.[4] This allows to estimate a detailed set of heterogeneous effects by the worker and firm characteristics that help to explain the underlying mechanisms of immigration adjustments. A critical characteristic of labor markets in developing countries is the interaction between the informal and formal sectors, especially in the smallest firms (Ulyssea, 2018). This connection is relevant in the face of immigration shocks, as Corbi et al. (2021) for Brazil, Delgado-Prieto (2022) for Colombia, and Kleemans and Magruder (2018) for Indonesia have found asymmetric employment and wage responses across these sectors. Specifically, Delgado-

[1] For example, Afghanistan, Ukraine, Syria, and Venezuela, among others.

[2] Throughout this paper, formal workers refer to workers who contribute monthly to the health system in Colombia.

[3] By following workers over time, I can address more carefully the compositional changes that arise in the regional-level analysis of immigration with cross-sectional data.

[4] In developed countries, Bratsberg and Raaum (2012) for Norway, Foged and Peri (2016) for Denmark, Dustmann et al. (2017) for Germany and Orefice and Peri (2020) for France have recently studied immigration shocks exploiting administrative data.

Prieto (2022) studies the Venezuelan immigration in Colombia to find a significant negative effect on regional formal employment, driven by small firms, that motivates the worker-level analysis –exploiting the firm dimension– done in this paper.[5]

The empirical strategy of this paper compares similar workers in areas with different exposure to migration over time. Because migrants endogenously sort into areas that offer the best economic opportunities for them, I use two distinct instruments: past settlements of Venezuelans and distance to the nearest crossing bridge with Venezuela. I exploit these instruments in a differences-in-differences research design (DiD-IV) to find a persistent negative impact on individual formal employment and wages for natives. The negative employment effect is driven by workers earning the minimum wage, those above 35 years old, and those who were self-employed before the immigration shock. A relatively high and binding minimum wage for many formal workers limits the space for downward wage adjustments (around 40% of all formal workers in Colombia earned the minimum wage in 2015) and increases their chances of job displacement.

On the other hand, the negative wage effect mainly affects native workers earning above the minimum wage and working in the smallest firms. The fact that migrants tend to concentrate in small firms, which are more constrained by the minimum wage and employ a higher share of informal workers, helps to rationalize this finding.[6] In terms of regional mobility, results suggest that formal workers are not systematically moving to other places after migrants arrive, although younger workers tend to be more mobile.

To rationalize how firms interact with immigration effects, I construct an imperfectly competitive labor market model based on Card et al. (2018) with heterogeneous firms, but adapting the labor inputs cost similarly to Ulyssea (2018) and allowing for imperfect substitution of labor inputs as in Delgado-Prieto (2022). The model shows that the substitutability between formal and informal workers must be high for a negative formal employment and wage response. The model also shows that firms are *more responsive* to an immigration shock depending on how much they initially rely on informal workers for production. Hence, the model predicts smaller firms react

[5] The focus of Delgado-Prieto (2022) is to understand the spillovers and interactions between the formal and informal sectors from an immigration shock, while this paper zooms into the formal sector to try to understand with more detail what happens to workers across different firms.

[6] For these reasons, the motivation to exploit the firm component in this context is different from the imperfect labor market argument in Amior and Stuhler (2022). Their idea is that the lower reservation wages of migrants have impacts across the distribution of firms, and this magnifies the negative employment effects of immigration beyond the predicted values in the competitive models.

more adversely to informal supply shocks than larger firms. Notably, these patterns are matched in the empirical findings, that is, workers in the smallest firms are more affected in terms of formal employment and formal wages than workers in larger firms.

Next, I exploit the canonical Abowd et al. (1999) (AKM hereafter) framework to recover firm fixed effects (FEs), firm-specific pay premiums, and worker FEs, or worker-specific pay premiums. A significant contribution of this paper is to understand the sources of wage and employment losses stemming from immigration using these constructed variables. I find that workers in middle-paying firms during the pre-shock period suffer the largest wage losses compared to workers in low-paying firms. A binding minimum wage in low-paying firms prevents wage losses, while for the rest of the firms, a potential explanation is the reallocation effects of immigration, that is, workers might be moving from middle- to low-paying firms. However, I find no differential sorting of native workers after migrants arrive. Hence, there seems to be lower wage growth within middle-paying firms after migrants arrive.[7]

Regarding employment, the finding is the opposite: native workers in low-paying firms have a more negative impact than workers in high-paying firms. A similar picture is found when dividing workers by their worker-specific pay premiums. Next, I analyze other firm outcomes changing in response to the immigration shock. Particularly, I show that firms opt-out from the formal sector for new hires, especially those more "connected" to the informal sector, and that the formal firm exit rate is higher in places that receive more migrants.

In the second part of this paper, I estimate the heterogeneity of treatment effects according to a vector of worker and firm characteristics following the recent literature in machine learning (Athey and Imbens, 2016; Athey et al., 2019). Specifically, I implement different causal forests that quantify a set of reduced-form estimates from random subsamples to determine those variables that explain most of the heterogeneity of immigration effects. From this algorithm, first, I identify the subgroups most affected by immigration, both for employment and wages. Then, based on the frequency that these variables appear in the splits of all the decision trees, I construct a simple measure of variable importance. In this exercise, I consistently find that firm-specific pay premiums are ranked higher in the algorithm, meaning they are more likely to explain employment and wage losses than other

[7] Theoretically, according to Amior and Stuhler (2022), when the share of firms in the low-pay sector gets bigger due to immigrants, firms in the high-pay sector increase their monopsony power and reduce workers' wages.

worker characteristics (i.e., job tenure, age, sex, and wages) in the pre-shock period. Therefore, firms' role in the impact of immigration appears to be very relevant, which is one of the main findings of this paper. These results suggest direct policy implications. For instance, to alleviate the largest losses of immigration, specific firm-level policies might be determined to enhance the productivity of the smallest firms so that employment for natives in the formal sector is not diminished.

Literature. Two of the first papers that estimated the effects of immigration at the worker level are Bratsberg and Raaum (2012) and Foged and Peri (2016). The first paper uses licensing requirements in the Norwegian construction sector to leverage exogenous immigration shares. These authors find that native wages in this sector decrease as immigrant shares increase, with low-paid native workers leaving this sector more frequently. The second paper exploits a refugee dispersal policy in Denmark to find that low-skilled natives pursue less manual-intensive occupations, upgrading their wages. A third paper by Dustmann et al. (2023) discusses the differing labor market effects of immigration at the region- and worker-level exploiting a commuting policy in Germany. They show that the region-level and worker-level effects measure conceptually different types of labor market responses to immigration. In the US, Brinatti et al. (2023) measure the impact of the H-1B visa lottery on firms and workers using administrative panel data, they find that affected firms grow in scale and productivity.

Other papers exploit longitudinal administrative data to study questions related to migration. For instance, Orefice and Peri (2020) uses a matched employee-employer dataset of France to study the impact of immigration on worker-firm sorting. This paper finds an increase in assortative matching, with high-quality firms attracting high-quality workers after migrants arrive. Another related paper by Arellano-Bover and San (2020) studies the role of firms in the assimilation of immigrants in the labor market. Exploiting the arrivals of former Soviet Union Jews to Israel in the 1990s and an augmented AKM specification, these authors find that firm-specific pay premiums explain around one-fifth of the immigrant-native earnings gap, as in Canada (Dostie et al., 2021).

In the trade literature, Autor et al. (2014) have estimated worker-level effects in the US using industry shocks to import competition from China. This paper finds a negative impact on earnings, especially for low-wage workers with lower job tenure. From labor demand shocks, Yagan (2019) uses individual-level data of workers in the US to analyze the impact of local unemployment shocks arising from the Great Recession. The main finding is that the Great Recession generated permanent employment and wage losses. Last, Gulyas et al. (2019) exploits mass layoffs in Austria, combined with a machine learning algorithm, to find that cumulative earning losses on workers are higher among the ones employed in high-paying firms.[8] Summarizing, all these papers only focus on workers in developed countries, while for developing countries, the literature is much more scarce regarding worker-level effects.

This paper contributes in different ways to the migration literature. The first contribution is to the impact of immigration in developing countries.[9] Since having administrative data in developing countries for workers and firms is unusual, most previous studies used cross-sectional surveys to determine the impact of immigration. Therefore, with panel administrative data, I quantify for the first time in a developing country the displacement effect on natives and the change in the price of labor across the distribution of firms. The second one is identifying the main drivers behind labor market adjustments to immigration. In this respect, I show that firms play a significant role in determining the wage and employment losses of immigration on native workers, both theoretically and empirically. This important result speaks directly to the literature on the labor market impacts of immigration that focuses mostly on workers' characteristics to understand the adjustments to labor supply shocks. The third contribution is to estimate the individual impact of immigration. With the universe of formal workers, I integrate into the analysis all the movements of natives between areas, reducing the attenuation of the wage estimates discussed in Borjas (2006), and exclude all inflows into employment that can affect the regional-level responses.[10]

The rest of the paper is structured as follows. Section 2 discusses the characteristics of the labor supply shock derived from the Venezuelan crisis and describes the data used. Section 3 describes the empirical strategy and the required identification assumptions. Section 4 reports results at the worker level by different individual characteristics. Section 5 introduces a model with heterogeneous firms and shows results by firm characteristics. Section 6 introduces the machine learning approach

[8]Relatedly, Yakymovych et al. (2022) uses a generalized random forest, with administrative data from Sweden, to identify sets of workers more vulnerable to job displacement and to uncover the subgroups with the most significant earnings losses of displacement. They find that older and less-educated workers are the most affected.

[9]See related papers of Morales-Zurita et al. (2020), Caruso et al. (2021), Lebow (2021), and Delgado-Prieto (2022) for Colombia; Del Carpio and Wagner (2015), Ceritoglu et al. (2017), and Aksu et al. (2018) for Turkey; and Groeger et al. (2022) for Peru.

[10]Recent regional-level studies are Monras (2020) in the US and Muñoz (2021) in the EU. The first found that low-skilled Mexicans who left their country due to the peso crisis had a high transitory impact on local labor markets in the US. The second exploits a trade liberalization in services across Europe to find a negative regional effect on the employment of domestic workers.

used in the paper and discusses the main findings of this method. Section 7 provides robustness tests. Finally, Section 8 concludes.

2 Institutional Context and Data

2.1 The Venezuelan Mass Migration

Historically, Colombia and Venezuela shared an extensive territorial border characterized by a dynamic relationship with frequent economic and cultural interactions. People often moved back and forth between the two countries, although many Colombians settled in Venezuela. This trend intensified after 1950, fueled by the oil boom in Venezuela and the internal conflict in Colombia. The economic opportunities presented by Venezuela's oil industry attracted many Colombians to emigrate, seeking better livelihoods and prospects for their families. Recently, the trend reversed with Venezuela's unprecedented socio-economic and political deterioration that triggered massive outflows of people since 2013, both voluntarily and forcedly. As a result, several countries in Latin America are receiving vast numbers of migrants, especially Colombia, Perú, and Ecuador (UNHCR, 2019). By far, Colombia has been the major receiver country with more than 1 million working-age Venezuelans (4.1% of the working-age population living in Colombia) as of 2019 (DANE, 2019). These sudden inflows alter different socio-economic outcomes in the short and long run in the host countries.

The Venezuelan exodus is unmatched in the recent migration history in Latin America. Worldwide, there are only two contexts with similar numbers, namely, the Syrian and the Ukrainian exodus. In the first case, Turkey has been the major receiver country of Syrians, with various papers analyzing this labor supply shock (e.g., Del Carpio and Wagner (2015); Ceritoglu et al. (2017); Aksu et al. (2018)). However, there are different characteristics between the Colombian setting and the Turkish one. First, Venezuelans speak the same language as Colombians, and second, Colombia's government has implemented an open border policy in which all Venezuelan immigrants may get work permits. In particular, after 2017, all undocumented Venezuelans in Colombia have access to the Special Permit of Permanence (PEP, by its acronym in Spanish). This allows them to work for a specific period, provides access to basic services, and facilitates their integration into Colombian society.[11] Yet, around 90% of Venezuelan immigrants were employed in the informal sector in 2019 –meaning they do not contribute to the social security system– and were concentrated at the bottom of the native wage distribution (Delgado-Prieto, 2022). This fact relates to the occupational downgrading of Venezuelans since they have similar average levels of education compared to their Colombian counterparts and are even more educated in the latest years of arrival.

As described above, the labor supply shock in Colombia occurs mainly in the informal sector, so why focus on the formal sector in this paper? Theoretically, I develop a model of heterogeneous firms to show that wages and employment in the formal sector are directly affected in response to the informal supply shock. As firms combine formal and informal labor in their production function (especially the smallest firms), they can substitute formal for informal workers in response to shocks when both types of labor have a high degree of substitutability. Therefore, native formal employment is the most affected, even if migrants primarily work in the informal sector. Focusing on formal workers' adjustments across the distribution of firms, therefore, is of central interest.

2.2 Data

The administrative data source for Colombia is the *Planilla Integrada de Liquidación de Aportes* (PILA), which contains administrative records from the Colombian social security system managed by the Health Ministry (*Ministerio de Salud y Protección Social*). PILA contains information on the universe of formal workers in tax-registered firms. It excludes informal workers and informal firms but includes self-employed formal workers. The PILA is based on the monthly contribution of the worker, according to their reported base income, to the health system in Colombia. Each observation in PILA is a worker-employer match for a given year and month. The dataset contains worker-level information on labor income, sex, age, job type (employee or self-employed), foreign status, municipality, and the firm identifier for each job. I have access to PILA from 2010 to 2019 for the month of August.[12] In addition, I use the most recent Colombian census (CNPV, by its acronym in Spanish), recollected between January and October of 2018, to construct the

[11] To overcome the limitations of the PEP, the government enacted in 2021 a Temporary Protection Statute for Migrants (ETPV, by its acronym in Spanish) that grants up to ten years of regularization for Venezuelan immigrants.

[12] The selection of August is due to the seasonal characteristics of other months (e.g., December-January or March-April) and because the census recollection ended in October of 2018, omitting arrivals of migrants in November and December of that year.

immigration shock. The census provides the most reliable source of information on Venezuelan immigrants.[13]

For the analysis, I built a dataset with all the individuals that appeared in PILA, from 2012 to 2019, in the rows and their yearly variables on the columns.[14] The total number of workers in this dataset, who appeared at least one year, is 18,430,987. Next, I restricted to full-time native workers between 25 and 55 years in 2015 and assigned the immigration shock to all these workers according to their 2015 location, which leaves 7,123,223 workers.[15] Then, I transform the municipality variable into a more standard definition of local labor markets or commuting zones, adjusting the methodology of Sanchez-Serra (2016).[16] This adjusted definition yields 113 functional urban areas (FUAs) after eliminating small or rural municipalities, with a sample of 6,706,035 workers.[17] This is the sample used for the employment analysis over all the years (a balanced panel). For the wage analysis, I further restrict to workers with 30 days of employment in the month and positive wages; moreover, the worker must be employed in the post-treatment year of comparison. Thus, the sample varies slightly by year (an unbalanced panel). For the impact of immigration on firm characteristics, I eliminate self-employed workers from the sample as they do not belong to a comparable firm definition. It is worth noting that all the restrictions to the dataset are the typical ones implemented in the literature.

Descriptive Statistics for Formal Workers. Table 1a, 1c and 1b shows descriptive statistics for natives, foreigners, and Venezuelans with PEP by age, sex, and wages across time.[18] In terms of

[13]The labor force survey (GEIH, by its acronym in Spanish) also measures the number of Venezuelan immigrants in Colombia at a higher frequency. Nevertheless, Aydemir and Borjas (2011) document a measurement error in surveys that might attenuate the immigration effects.

[14]The administrative records of PILA are constructed at the level of the contribution, as workers with more than one labor contract need to pay contributions for each one of them. To transform to a worker-level dataset, first, I drop out all the contributions to the health system with type N, which are the ones that present corrections to their base income or changes to their labor status. Second, I aggregate the income for workers with more than one labor contract and leave the characteristics only for the job with the highest base income by the worker.

[15]Selecting only the workers observed in the base period rules out inflows of workers in the post-treatment period from the analysis. Also, in 2015, part-time workers in PILA were very few.

[16]A shortcoming with any municipality variable in PILA is that certain firms with several establishments across the country report the information for all employed workers in the municipality where the biggest establishment is located.

[17]The definition of FUAs consists of the 53 most extensive urban areas in the country defined from population grid data, municipal boundaries, inter-municipal commuting flows, plus 56 municipalities with more than 2,500 formal workers according to the restricted sample in 2015. I exclude San Andrés, Cumaribo, Leticia, and San José del Guaviare from this definition as they belong to the islands or Amazonia. In Appendix Table G.1, I show the sample distribution by FUAs; using this definition, only 5.9% of workers are excluded.

[18]To identify foreign status in PILA, I exploit the type of document workers have in their health contribution. If workers have a national ID, they are defined as natives, whereas if their document refers to the Special Permit of Permanence (PEP, by its acronym in Spanish) they are defined as Venezuelan migrants. Since the PEP's program

observable characteristics, Venezuelans with PEP in the formal sector are younger, predominantly male, and earn lower wages compared to natives and other foreigners (see Table 1a, 1c and 1b). In fact, the group of foreigners earns substantially higher wages than natives. In addition, the share of Venezuelans with the PEP working in the formal sector is small, supporting the fact that the impact of the PEP regularization on the Colombian labor market is limited (Bahar et al., 2021).[19] Note that it is not possible to observe informal workers in the administrative data, but they represent around half of all workers employed in Colombia.

Table 1: **Descriptive statistics for natives and migrants by years in the formal sector**

(a) Colombians

	Age		Male (%)		Real wages (USD)		
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	N
2013	37.0	10.8	0.56	0.496	421.4	504.0	7,335,989
2015	37.2	11.1	0.56	0.497	416.8	484.5	8,391,843
2017	37.8	11.4	0.55	0.497	411.4	477.8	8,064,282
2019	38.2	11.7	0.55	0.498	436.1	505.7	8,363,249

(b) Venezuelans with PEP

	Age		Male (%)		Real wages (USD)		
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	N
2018	30.8	7.8	0.68	0.467	243.7	99.0	12,842
2019	31.8	8.1	0.67	0.472	248.7	98.5	42,752

(c) Other foreigners

	Age		Male (%)		Real wages (USD)		
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	N
2013	40.2	10	0.61	0.487	1,310.0	1,535.9	20,978
2015	39.5	10.2	0.64	0.481	1,253.4	1,446.6	27,730
2017	39.3	10.3	0.63	0.483	1,050.8	1,325.0	31,553
2019	39.6	10.4	0.58	0.494	990.5	1,293.0	39,704

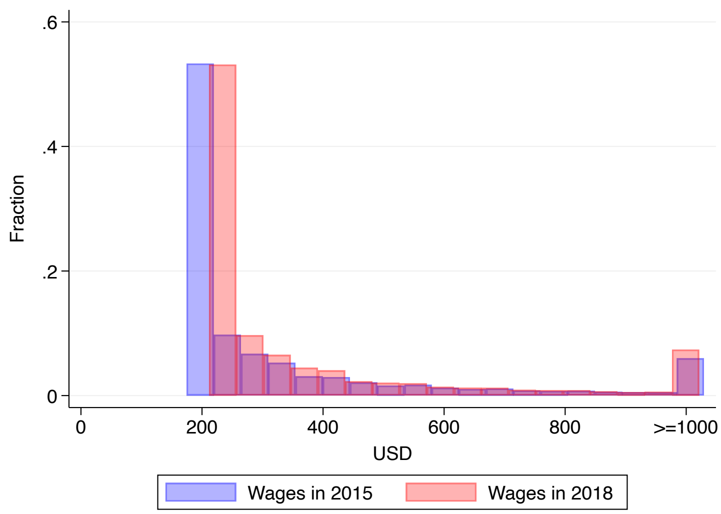
Note: This table reports the descriptive statistics for Colombians, foreigners, and Venezuelans with PEP between 18 and 64 years of age. Only workers with full days of employment recorded in PILA and a positive health contribution are considered for wages and the number of observations. I only observe Venezuelans with PEP since 2018, after the law's enactment. The real wages are deflated using the Consumer Price Index (CPI) from DANE for prices in 2018. Colombian pesos to USD using 2020 exchange rates from the World Bank. For self-employed workers, observed wages in PILA correspond to 40% or more of their actual wages by law, with the minimum wage as a lower bound. Source: PILA, 2013–2019, August.

started around 2018 to foster the regularization of Venezuelan migrants, it was not possible to identify these migrants before that year. Last, if their document is a foreigner ID or a passport, they are defined as foreigners.

[19] Note that other Venezuelans might be working in the formal sector in the group of foreigners, but it is not possible to identify them.

Figure 1 shows how binding the minimum wage in Colombia is for a large portion of formal workers. In 2015, around 40% of all formal workers earned the minimum wage.[20] Importantly, the national minimum wage in Colombia must increase yearly, by law, more than the inflation in the preceding year. This downward rigidity suggests why, in general, there are no real wage drops (but more layoffs) for minimum wage workers in the face of positive labor supply shocks or negative demand shocks. Last, in the period of analysis (2015-2109), the minimum wage increased in real terms by less than 3% each year, reducing the concern of additional impacts of the minimum wage on employment.

Figure 1: **Histogram of wages by years**



Note: The sample is restricted to native workers between 18 and 64 years with full employment days in the month and positive wages. Wages are in nominal terms. Colombian pesos to USD using 2020 exchange rates from the World Bank. Chosen bin width is 45. Source: PILA, 2015–2018.

Descriptive Statistics for Formal Firms. I aggregate the worker information of the PILA at the firm level to describe patterns in the workforce composition of formal firms.[21] Table 2 is split into seven firm size categories to show certain facts. First, regarding sex, male workers are the main group in all formal firm sizes, especially for small-medium firms (between 10 and 999 workers), where more than 60% of formal workers are males. Second, smaller firms have older workers on average (39.2 years), while larger firms have younger workers (35.6 years). Third, average wages are

[20]To contribute to the health and pension system in Colombia, the worker must declare a labor income equal or greater to the minimum wage, so many self-employed workers (who decide how much is their observed income in PILA) declare the minimum wage even if they earn more or less.

[21]For the firm analysis, I eliminate self-employed workers from the main sample.

growing with firm size, from around 272 USD in firms with 1 to 4 workers to 531 USD in firms with more than 1000 workers. Last, Appendix Figure C.1 plots the histogram of firms by size overlay with the total number of employees in each firm size. Although most firms are concentrated in the interval size of 1 to 4 employees, the number of employees is more evenly distributed across different sizes of firms.

Table 2: **Descriptive statistics by firm size**

	Average				
Firm size (# of workers)	Employment	Male (%)	Age	Real wages	N
1-4	2	0.56	39.6	271.9	206,456
5-9	7	0.60	37.5	304.2	64,347
10-19	13	0.61	37.2	329.2	42,207
20-49	30	0.63	36.9	360.9	28,625
50-99	69	0.65	36.7	394.8	10,032
100-999	259	0.63	36.9	443.1	10,107
1000 and more	2677	0.58	36.0	530.8	859

Note: This table reports the descriptive statistics for different firm sizes recorded in PILA. Real wages are deflated using the CPI from DANE for prices in 2018. Then, I transform Colombian pesos to USD using 2020 exchange rates from the World Bank. Only workers who contribute as employees are taken into account. Source: PILA, 2015-August.

3 Empirical Strategy

To quantify the evolving impact of immigration on individual wages, I estimate the following differences-in-differences (DiD) specification from $t = \{2012, .., 2019\}$ estimating separate regressions of the following form:

$$\frac{w_{ilt} - w_{il,2015}}{w_{il,2015}} = \delta_t + \theta_t \Delta M_{l,2018} + X_i' \beta + \Delta u_{ilt}. \tag{1}$$

Here, $w_{i,l,t}$ is the wage of worker i in labor market l in period t. The dependent variable measures the percentage change of wages for each worker with respect to 2015, and the vector X_i contains individual characteristics in 2015, namely, interactions of FEs for seven age groups with dummies for sex and self-employment.[22] Using this specification, I compare individuals who were employed in 2015 and the year of comparison t, with similar observable characteristics but working in different local labor markets in the face of the immigration shock $\Delta M_{l,2018}$, which I will describe

[22] Industry FEs are not used as controls because it is not possible to observe without measurement error that variable in PILA.

below in detail. Moreover, individual constant unobservable characteristics are ruled out from the analysis by taking differences. In this case, $\theta_{2015} = 0$ by construction.

Next, to quantify the evolving impact of immigration on individual employment, I estimate separate regressions, as in Yagan (2019), between $t = \{2012, .., 2019\}$ of the following form:

$$e_{ilt} - \sum_{k=2013}^{2015} e_{ilk}/3 = \omega_t + \gamma_t \Delta M_{l,2018} + X_i' \lambda + \Delta\epsilon_{ilt}, \tag{2}$$

where $e_{i,l,t}$ is the indicator of being employed in the formal sector for worker i in labor market l in period t. I consider the average employment in the pre-shock period to transparently allow for varying labor trajectories of workers in the formal sector. However, in the event study figures, I take the simple difference with the base period ($e_{ilt} - e_{il,2015}$) to avoid pre-treatment coefficients being mechanically around zero. The constant for each year is δ_t and ω_t, and the standard errors in all the specifications are clustered at the level of the treatment, which are the FUAs ($G = 109$).

The immigration shock $\Delta M_{l,2018}$ is defined as follows:

$$\Delta M_{l,2018} = \frac{L_{Ven,l,2018} - L_{Ven,l,2015}}{L_{Total,l,2018}}, \tag{3}$$

where the numerator is the stock of employed migrants from Venezuela (between 18 and 64 years) in local labor market l who arrived in Colombia in the previous 5 years, starting from 2018, minus the stock of employed migrants from Venezuela in l whose year of arrival was 2015 or earlier according to the census. Employed migrants are either Venezuelans or returning Colombians from Venezuela, and the denominator $L_{Total,d,2018}$ is the total employed population in the local labor market. I focus and interpret mainly the coefficient of 2018 in the regressions (i.e., θ_{2018} and γ_{2018}) to match the year of the census from which the immigration rate is constructed.

Because migrants self-select into areas where the economic opportunities are better, the immigration rate $\Delta M_{l,2018}$ is likely to be endogenous, and its coefficient is downward biased (see ordinary least squares (OLS) estimates of Figure 3a and 3b). Thus, to consistently estimate the effect of immigration on the outcome variables, the immigration rate $\Delta M_{l,2018}$ is instrumented with the distance to the nearest crossing bridge with Venezuela and with past settlements of Venezuelans. The motivation for the IV approach is the following.

First, distance is exploited as an instrument since Colombia and Venezuela share 2,220 kilometers of terrestrial borders. Therefore, arrivals to the local labor market l are a function of travel distance between the two countries, as distance acts as a time and economic constraint for Venezuelan immigrants. A threat to this identification strategy is that border departments might be more affected, in terms of economic shocks (such as less trade), than the counterpart far-located states from the Venezuelan crisis (violation of the exclusion restriction).

Figure C.2a shows suggestive evidence that the trade shock arising from the Venezuelan crisis started years earlier than the immigration shock. Importantly, in the post-treatment period, border department exports to Venezuela are regularly around zero. Another important piece of evidence is that I find insignificant employment and wage effects in the largest firms, presumably more affected by trade shocks and less affected by migration shocks (as migrants disproportionally concentrate in small firms). In addition, I plot log GDP for border and non-border departments over time to show that it is evolving similarly before and even after the immigration shock, suggesting that trade's impact on economic activity is limited (see Figure C.2b). Last, I exclude border areas from the main sample and find similar point estimates, only not significant for wages. With this suggestive evidence in mind, formally, it is required that distance fulfills the following exogeneity assumption $E[f(dist_l)\Delta u_{lt}] = 0$.

The other instrument constructed uses past settlements of Venezuelans in the spirit of Altonji and Card (1991) and Card (2001), and it is defined as:

$$z_l = \left(\frac{Ven_{l,2005}}{Ven_{2005}} * M_{2018}\right)/L_{l,2005}, \tag{4}$$

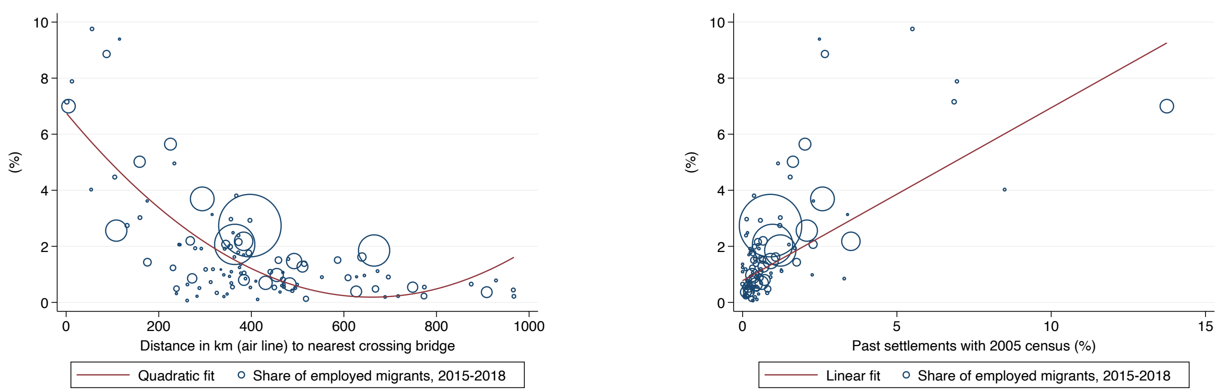
where the first term is the share of Venezuelans in FUA l (according to the 2005 population census), normalized by the working-age population $L_{l,2005}$ in l at 2005, whereas M_{2018} is the number of migrants in Colombia that arrived between 2018 and 2016 according to the census. I use past settlements as the other instrument because newly arriving immigrants will likely move to areas with previously established Venezuelans. To have a valid instrument, it is required that past settlements are related to new arrivals but not related to time-varying shocks (i.e., $E[z_l\Delta u_{lt}] = 0$).

Figures 2a and 2b show the first stage of the migration shock $\Delta M_{l,2018}$, for the 109 FUAs defined for this analysis, against the instruments. I show the instruments' relevance and functional in these figures. For the first instrument, a larger distance from a crossing bridge decreases the share of employed migrants in the FUAs until a point where longer distances do not imply lower immigration rates, that is, the slope of the curve bends downward. For past settlements, there is a positive relationship against the immigration rate that appears to be linear. The immigration shock at the FUA level is quite large, some experience an increase in the share of employed migrants that represent between 7% and 10% of their overall employed population.[23]

Figure 2: **Immigration rates and the two instruments**

(a) Distance

(b) Past settlements



Note: Dots are weighted by formal employment according to the PILA. Functional Urban Areas in Colombia (G=109). Source: CNPV, 2018.

Figures 2a and 2b are constructed at the FUA level. Yet, since this paper aims to estimate the impact of immigration at the individual level, the first stage of the two-stage least squares regression (TSLS) is going to weigh each FUA differently by the number of individual observations available.[24] With this in mind, the first-stage model is:

$$\Delta M_{l,2018} = \delta + f(dist_l) + z_l + \upsilon_l \tag{5}$$

Here, $f(dist_l)$ is equal to the two polynomials of distance to the nearest crossing bridge, whereas z_l are the past settlements of Venezuelans. In this equation, the error term is υ_l, which captures the endogenous component of $\Delta M_{l,2018}$. I combine the two instruments in the analysis as past

[23] In Delgado-Prieto (2022), the department is used as the area of analysis because of sample limitations of the GEIH survey, but with administrative data, there are no sample issues when constructing more detailed areas.

[24] Hence, the first stage varies slightly depending on the sample used.

settlements or distance capture different exogenous components of migration while increasing the R^2 of the first-stage regression (see Table A.1).[25] As a result, equations (1) and (2) are estimated throughout the paper using TSLS with the aforementioned instruments.

4 Worker-level Effects

This section documents the impact of immigration on formal wages and employment at the worker level. First, I show wage and employment estimates using OLS and TSLS. One advantage of the proposed empirical specification is that it is possible to test for differential trends of the outcome before the immigration shock happens. Importantly, there are no significant pre-trends for employment and wages that can confound the impact of immigration. Yet, the OLS coefficients are close to zero, presumably downward biased, as immigrants are expected to arrive in those areas with better economic opportunities. Hence, one should expect the TSLS coefficient to become more negative. Specifically, for 2018 I find that a one percentage point (pp) increase in the share of employed migrants in a given area reduces the probability of being employed in the formal sector by –1.1 pp (see Figure 3a).[26] This means that the probability decreases by 2.4% for every pp increase in the share when using the labor force survey to measure the probability of being employed in the formal sector.[27] More broadly, a worker located in the LLM in the 75th percentile of exposure relative to one in the 25th percentile of exposure is around 3.6% less likely to remain employed formally.[28] Regarding formal wages, I find a coefficient of –0.6% in 2018 for a one pp increase in the immigration shock (see Figure 3b).

[25] Moreover, main results do not change significantly if I use one of the instruments instead of the two.

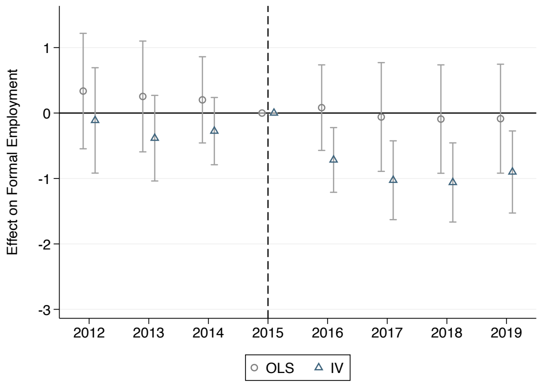
[26] This regression uses as dependent variable $e_{i,l,2018} - e_{i,l,2015}$, which captures the difference in the employment indicator in 2018 relative to the base period. In the heterogeneity analysis, the regression uses as a dependent variable the employment change $e_{i,l,t} - \sum_{k=2013}^{2015} e_{i,l,k}/3$, which yields slightly less negative coefficients as it uses for comparison the average of employment indicators in the pre-shock period.

[27] In the survey, I restrict to workers between 25 and 55 in 2015, which yields a probability equal to 0.42.

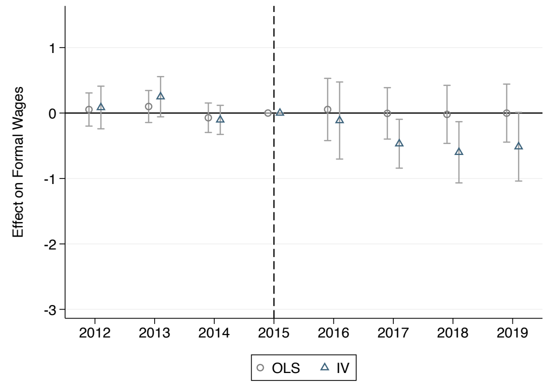
[28] The 75th and 25th percentile migration rate is 2.1% and 0.6%, respectively. So, (2.1-0.6)*2.4=3.6.

Figure 3: **Event study estimates on individual wages and employment**

(a) Employment



(b) Wages



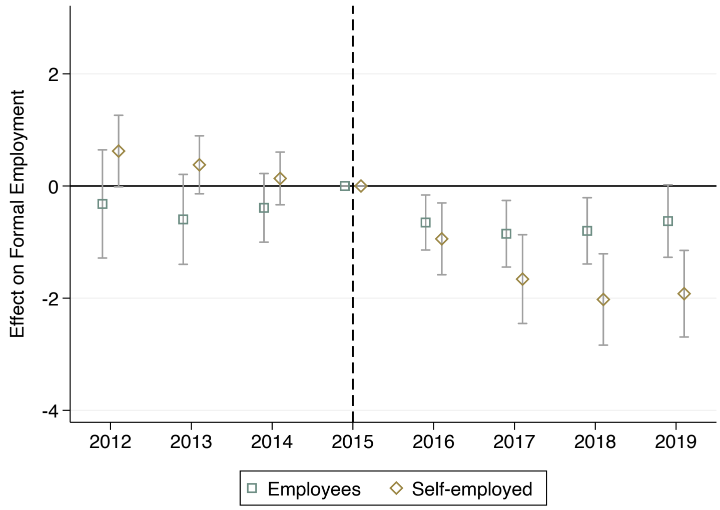
Note: The sample in panel (a) is 6,706,035 workers, while in panel (b), it varies slightly by year as the workers must be employed in the post-treatment and base year. Moreover, the sample is restricted to natives between 25 and 55 years old. I use as controls interactions of sex with six age categories and a dummy for self-employed in the base period. I cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA 2013–2019.

For the rest of the paper, I focus on wage and employment estimates using the characteristics of workers before the immigration shock, specifically, the characteristics in 2015. In this case, the coefficients for each subgroup come from separate regressions of the main empirical specification. The first worker observable is by job type, there are two categories: employee and self-employed.[29] Self-employment in Colombia represents about half of the employed population, mainly working in the informal sector but with a large share of workers in the formal sector (around 18% of all native formal workers were self-employed in 2015). Figure 4 shows a drop in the probability of being a formal worker for self-employed natives, which is more negative than native employees. Most self-employed in the private sector decide voluntarily whether to contribute or not to the social security system, so opting out from the formal sector is less costly for them than for employees.[30]

[29] I use only IV hereafter because OLS estimates are inconsistent (see Figures 3a and 3b).

[30] The labor income for self-employed is noisy in PILA, but the point-estimates on wages are also more negative than for employees.

Figure 4: **Event study estimates on employment by job type**



Note: The sample is 6,706,035 workers and is restricted to natives between 25 and 55 years old. I use as controls interactions of sex with six age categories and a dummy for self-employed in the base period. I cluster standard errors (G=109). The coefficients for employment (in percentage points) are already multiplied by 100. Workers are observed in August of each year. Source: PILA 2013–2019.

After showing suggestive evidence that the instruments do not predict native wages or employment trends in the pre-treatment period, I focus, for the rest of the analysis, on the coefficient of 2018 (the year of the immigration shock from the census) exploiting certain characteristics of workers in 2015 (the year before the immigration shock).[31] I start this heterogeneity analysis with the standard variables used in the migration literature, but later on, I exploit firm characteristics and develop a more systematic heterogeneity analysis using a machine learning algorithm.

The first result is by age groups and sex, which are the controls used in the main specification. Figure 5a shows a decreasing pattern in the probability of being employed in the formal sector as the worker ages. In contrast, the pattern is not equally clear for wages, and I find similar negative estimates in all age groups. I extend the sample of analysis in Appendix Figure B.1 to include labor market entrants (18 to 24 years) and workers close to retirement (56 to 64 years) in the base period. For employment, the highest negative effect is observed in the oldest workers, suggesting that they are retiring earlier, while for wages, again, there are no stark differences. Last, in terms of sex, the impact on employment and wages is alike, there are no differential effects in this group category.

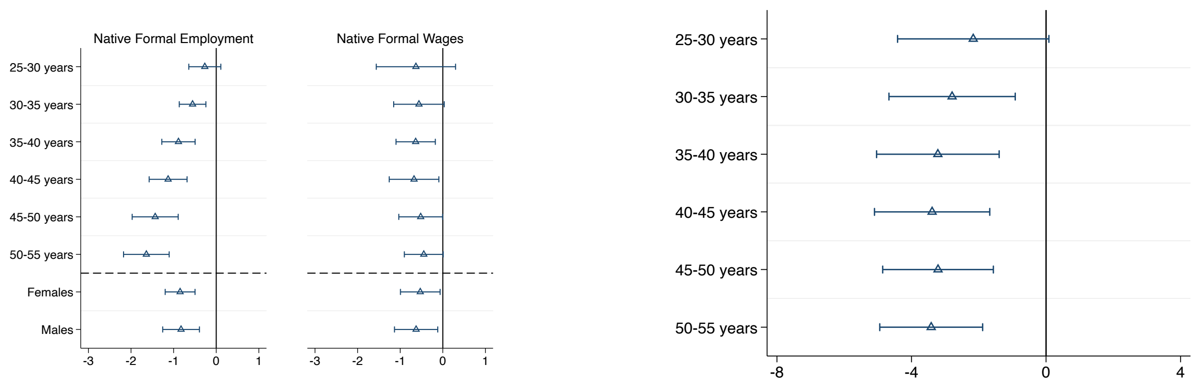
[31]Nonetheless, in Appendix Table F.1 I show there are no systematic pre-trends by certain categories on employment or wages.

Next, I construct a cumulative earnings outcome based on Autor et al. (2014). This variable is defined as $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,2015}}$ and it measures changes in the evolution of earnings normalized by the earnings in the pre-shock period. If the worker is not employed in the formal sector in any given period the earnings are zero, so this outcome yields a combined effect of the observed changes in employment and wages.[32] Figure 5b shows that workers above 30 years old present relatively a similar reduction in cumulative earnings comparing their confidence intervals. Indicating that even if older workers are more displaced from the formal sector, younger workers had a persistent reduction in their wages.

Figure 5: **Estimates by age group, 2015–2018**

(a) Employment and wages

(b) Cumulative earnings



Note: In (a), dependent variables are employment and wages relative to the base period. In (b) dependent variable is $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,2015}}$. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019 .

To complement the pattern of employment effects by age group, I also calculate the labor supply elasticities, at the extensive margin, for each of these age groups (i.e., $\eta_w^s = \frac{\Delta L}{\Delta w}$). Table 3 shows that as native workers age, their labor supply is more elastic. That is, the responsiveness to work from wage changes is more important for older than younger workers. Similar to Dustmann et al. (2017) findings for Germany, where the elasticity of labor supply is increasing in workers' age.

[32] Workers with less than 30 days of employment in the social security contribution are defined as workers with missing wages (the wage analysis is focused only on full-time workers).

Table 3: **Labor supply elasticities by age group**

Age group	25-30	30-35	35-40	40-45	45-50	50-55
η_w^s	0.42	0.99	1.39	1.67	2.76	3.67

Note: The elasticity of labor supply is given by the reduced-form results from changes in native employment over changes in native wages.

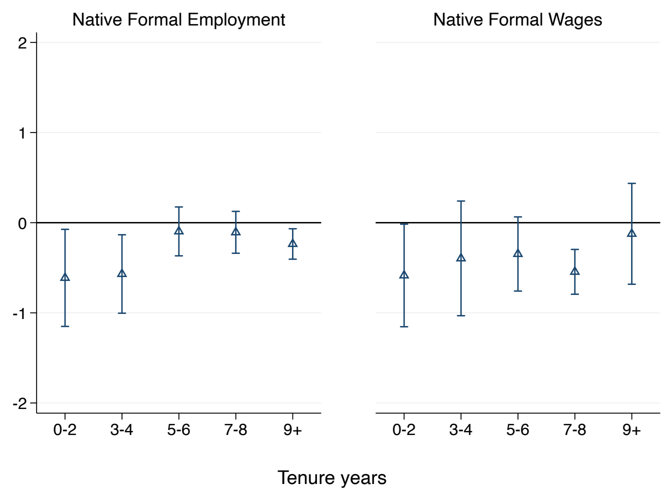
To understand in more detail which is the most affected type of worker, I exploit the number of years the worker has been employed in the same firm (i.e., job tenure) up to the base period 2015.[33] Figure 6 splits the sample by job tenure of native workers (from zero to more than nine years of tenure). Interestingly, the shock of employment and wages due to immigration is more severe on workers with fewer years in the same firm.[34] The pattern is more precise on employment than on wages. This result is partly explained by the fact that the severance payment increases with workers' tenure, so it is more costly for firms to dismiss workers, and partly by the accumulation of firm-specific human capital, as they are less substitutable to migrants with similar characteristics.

The last two figures suggest that older workers and workers with lower tenure have the most significant drop in formal employment from the immigration shock. To better explain how workers react, I combine their age and job tenure. Appendix Table B.1 shows that the age variable is more relevant for employment, as native workers below 35 present an insignificant effect on employment, independent of whether they have low or high job tenure. On the other hand, native workers above 35 present a significant negative effect on employment when they have low and high job tenure, but the effect is much higher for the workers with lower tenure (–1 pp versus –0.3 pp). Regarding wages, only younger workers with high tenure present a significant negative effect. More concretely, a 1 pp increase in the immigration shock reduces their wages by 0.7%.

[33]Self-employed workers are excluded from this analysis as they are not comparable to the average firm.

[34]I construct job tenure from the first year PILA is available (2007).

Figure 6: **Estimates by job tenure, 2015–2018**



Note: Dependent variables are employment and wages relative to the base period. The sample is restricted to native employees between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019.

4.1 Distributional Impacts of Immigration

I then estimate the impact of immigration on workers across the distribution of wages. All native workers employed in 2015 are divided according to their local wage distribution in that year. Figure 7a plots the coefficients for 2018 relative to 2015 according to this division.[35] It is remarkable the uneven impact of immigration: native workers earning the minimum wage suffer the most negative shock on formal employment, while for workers at the rest of the wage distribution, I find insignificant estimates on employment. For these low-wage workers, a one pp increase in the share of employed migrants in a given labor market reduces the probability of being employed in the formal sector by –1.5 pp. Interestingly, formal workers who earn the minimum wage are the least affected by the immigration shock in terms of wages. Because the minimum wage is relatively high and binding for around 40% of formal workers in the pre-shock period, there are no further drops in formal wages, and this muted wage effect increases the chances of job displacement.[36] Definitely, movements to the informal sector of these low-wage workers, which are not observed,

[35] As a robustness check, Appendix Table F.2 shows the pre-treatment coefficients by wage categories on employment and wages. Reassuringly, most of these coefficients are insignificant.

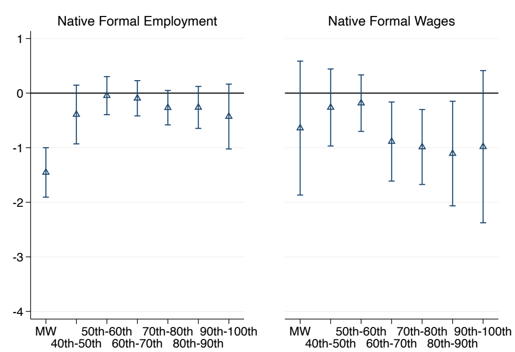
[36] Conditional on being employed in the two periods, around 75% of minimum wage earners still earn the minimum wage after three years.

explain part of the large coefficient, as minimum-wage workers are the less educated and, thus, the most substitutable with informal workers.

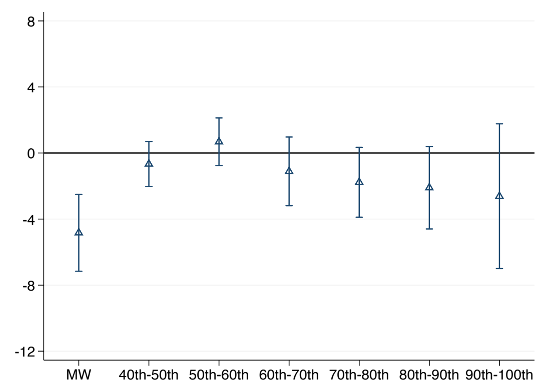
For the workers between the 60th and 90th percentile of the local wage distribution, I find a drop of around –1% and –1.2%. This does not necessarily mean a decrease in absolute terms of wages. The coefficient measures the average growth of wages of native workers in areas with more exposure to migration with areas with less exposure. A negative estimate means that the growth rate of wages in more affected areas is relatively lower.[37] Last, I include the cumulative earnings outcome to find which set of workers gets more affected overall. Interestingly, Figure 7b shows that the only significant negative impact on earnings is observed on workers who earn the minimum wage before immigrants arrive, reflecting a stronger effect from the employment margin.

Figure 7: **Estimates by individual wage at baseline, 2015–2018**

(a) Employment and wages



(b) Cumulative earnings



Note: In (a), dependent variables are employment relative to the pre-shock period and wages relative to the base period. In (b) dependent variable is $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,2015}}$. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019 .

4.2 Worker-level and Regional-level Estimates

To compare worker-level estimates with regional-level estimates of immigration, first, I adapt to this setup the employment decomposition formalized in Dustmann et al. (2023). Thus, I decompose the changes in regional formal employment into 1) a displacement of incumbent workers –outflows

[37]The exit of low-wage workers from formal employment plus an average drop of wages from workers in the middle and upper part of the wage distribution rationalizes the insignificant effect on regional formal wages of Delgado-Prieto (2022).

from formal employment–, 2) hiring of new formal workers or inflows from other regions –inflows to formal employment–, and 3) relocation of existing employed formal workers to other regions. In this analysis, the worker-level estimates of employment capture the outflows or the displacement of incumbent native workers in the formal sector, while the regional-level estimate from cross-sectional data combines the three margins of adjustment. Appendix Figure B.2 shows the decomposition of the regional formal employment response at the FUA-level (–1.3%) along with the three margins: outflows to non-employment or the informal sector (1.1%), inflows from other regions, non-employment or the informal sector (–0.5%) and relocation to other regions (–0.4%). In this case, the most important and only significant margin is the outflows from the formal sector, which differs from the findings of Dustmann et al. (2023), where inflows are the most relevant margin.[38]

Regarding wage estimates, the worker-level response is –0.6%, while the regional-level estimate in Delgado-Prieto (2022) is insignificant and close to zero. These two responses are complementary and answer different policy questions. As stated in Dustmann et al. (2023), the wage estimates of the worker-level regressions capture the change in the price of labor, holding the composition of the population constant, while the regional-level regressions jointly measure the change in the selection and composition of workers and the price of labor. The differential estimate between the two is rationalized in this setup as follows. The immigration shock changes the composition of employed natives and positively selects the individuals remaining in the region (see Employment of Figure 7a), therefore mechanically increasing regional formal wages. On the other hand, immigration decreases the price of labor in certain high-wage subgroups (see Wages of Figure 7a), reducing regional formal wages. Hence, this suggests why there is an insignificant wage effect at the regional level while having a negative wage effect at the worker level, motivating the analysis of immigration not only for the aggregate local labor markets but for each individual within local labor markets.

Another benefit of individual data compared to regional data is the possibility of estimating inter-regional movements of different types of workers to respond to the immigration shock. For instance, Dustmann et al. (2017) finds that younger workers are more mobile after immigrants

[38] The decomposition proposed by Dustmann et al. (2023) is equal to: $\frac{E_{r1}-E_{r0}}{E_{r0}} = -\frac{E_{r,Out}}{E_{r0}} + \frac{E_{r,In}}{E_{r0}} - \frac{E_{r,Move}}{E_{r0}}$. The first term measures the outflow margin, the second term measure the inflows margin and the third term measures the relocation margin. The main distinction, in this case, is that the outflows and inflows margins are decomposed further into non-employment or the informal sector. Unfortunately, there is no data to show this distinction.

arrive. Hence, Table 4 shows the impact of movements across regions by age groups. Indeed, younger formal workers tend to move more, but the coefficients are insignificant. Overall, the point estimates decrease as the worker ages, but all are insignificant. The mobility margin of adjustment is less important in this setup.

Table 4: **IV estimates on regional changes of formal workers by age group, 2015–2018**

Age group	25-30	30-35	35-40	40-45	45-50	50-55
Prob. of changing region	0.200	0.088	-0.035	-0.156	-0.211	-0.254
	(0.400)	(0.404)	(0.354)	(0.307)	(0.266)	(0.209)
N	1,255,301	1,041,726	873,437	732,208	674,945	561,949
Clusters	109	109	109	109	109	109

Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The outcome variable is an indicator that takes value one for workers that changed region in 2018 relative to 2015, and zero otherwise. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. The PILA had a measurement error with the regional variable in 2018, so the worker's location in February 2020 (when the health ministry started to verify this information) is used for the workers that present this error. Workers are observed in August of each year. Source: PILA, 2015–2018.

5 Immigration, Workers, and Firms

In this section, I first develop a partial equilibrium model with heterogeneous firms and types of workers to motivate and interpret the empirical findings. Then, I show that immigration effects for natives change substantially depending on the firm the worker was employed by before immigrants arrived.

5.1 Model

The market structure of the model consists of J firms that hire two types of labor inputs. Specifically, firms hire formal workers F paying payroll taxes and informal workers I off the books to avoid paying the payroll taxes, as in Ulyssea (2018). So, each firm $j = \{1, ..., J\}$ posts a pair of wages (w_{I_j}, w_{F_j}) that all workers i observe and decide to accept.[39] Importantly, each firm has different work environments, measured by amenities a_{L_j}, workers have idiosyncratic preferences ϵ_{i,L_j} depending on the fixed labor group they belong $L \in \{I, F\}$. This gives a different job valuation at

[39]The transitions of workers between the formal and informal sectors are out of the scope of the model.

each firm.[40] In this case, the indirect utility of worker i employed at firm j is:

$$\upsilon_{i,L_j} = \beta_L lnw_{L_j} + a_{L_j} + \epsilon_{i,L_j}. \tag{6}$$

Under the assumption that ϵ_{i,L_j} follows a type I extreme value distribution for each of the workers' types $L \in \{I, F\}$ and that the number of firms J is sufficiently large, Card et al. (2018) shows that the firm-specific supply functions is expressed as:

$$lnI_j(w_{I_j}) = ln(\mathcal{I}\lambda_I) + \beta_I lnw_{I_j} + a_{I_j}, \tag{7}$$

$$lnF_j(w_{F_j}) = ln(\mathcal{F}\lambda_F) + \beta_F lnw_{F_j} + a_{F_j}. \tag{8}$$

In this case, the total number of informal workers in the market is $\mathcal{I}$ and of formal workers is $\mathcal{F}$, where λ_I and λ_F are constant parameters across firms. Moreover, $\frac{dlnL(w_{L_j})}{dlnw_{L_j}} = \beta_L$ is the elasticity of labor supply to a given firm with respect to its wage. Hence, as $\beta_L \to \infty$, the supply functions become perfectly elastic, and firms have no monopsony power to set wages.[41]

Regarding firms, there is a productivity shifter T_j, a price of the good P_j, and a production function Q_j for each firm, such that the profit function of firm j is:

$$\max_{I_j,F_j} \pi_j = P_j T_j Q_j - \tau(I_j) w_{I_j}(I_j) I_j - (1 + \tau_F) w_{F_j}(F_j) F_j. \tag{9}$$

Here, $\tau(I_j)$ represents the standard convex cost increasing on the firm's informal labor size. Following Delgado-Prieto (2022), I assume that $\tau(I_j) = I_j^{\eta}$ with $\eta \geq 0$, which captures the cost of evasion related to law enforcement exerted by the government. The τ_F represents the payroll taxes firms must enroll in when paying for formal workers. In this setup, the production function takes the following form: $Q_j = (\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho})^{\frac{1}{\rho}}$. Thus, formal and informal workers are imperfect substitutes, and the aggregate elasticity of substitution common across all firm types is given by $\sigma = \frac{1}{1-p}$. To finish the setup, P_j is the inverse demand function defined as $P_j = D_j(T_j Q_j)^{-(1-\epsilon)}$,

[40] For instance, preferences for working in a firm may refer to distance to the workplace or interactions with coworkers (Card et al., 2018).

[41] Here, the labor supply function is constant in terms of β_L, and it does not include the reference market wage level for each labor group for simplicity. Moreover, the comparative statics focus on firm-level responses to immigration and not on market-level responses that have been thoroughly analyzed in previous studies.

where $\epsilon^D = -1/(1-\epsilon)$ is the elasticity of product demand and D_j is the firm-specific product demand.[42]

After describing the model, I analyze the impact of an immigration shock that shifts the aggregate informal supply outwards ($d\mathcal{I}$).[43] I study the firms' response across the wage and employment margin, so the wage elasticity for each type of worker in firm j is $\varepsilon_{w_{L_j},\mathcal{I}}$ and the employment elasticity for each type of worker in firm j is $\varepsilon_{L_j,\mathcal{I}}$. Allowing for distinct firm responses to an immigration shock is the main contribution of this framework. After some derivations in Appendix E, I show that the change in formal wages for firm j to an aggregate informal supply shock is equal to:

$$\varepsilon_{w_{F_j},\mathcal{I}} = \Omega_j s_{I_j}(\epsilon - \rho). \tag{10}$$

Here, $s_{I_j} = \frac{\alpha_I I_j^\rho}{\alpha_I I_j^\rho + \alpha_F F_j^\rho}$ is the relative contribution of informal work to production before immigrants arrive and $\Omega_j = \frac{1}{\xi_{I_j}\xi_{F_j} - (\epsilon-\rho)^2 s_{I_j}\beta_I s_{F_j}\beta_F}$ is a positive parameter.[44] The first implication is that formal wages in firms without informal workers do not change in response to an informal supply shock. In firms with informal workers, if they are close substitutes to formal workers (such that $\rho > \epsilon$), then the elasticity of formal wages will be negative. Importantly, as the contribution of informal labor to production in firm j increases ($s_{I_j} \uparrow$) the elasticity of formal wages is going to be more negative ($\varepsilon_{w_{F_j},\mathcal{I}} \downarrow$). Note that, for certain low productivity firms, formal wages might be downwardly rigid due to the existence of a minimum wage, so the formal wage margin might be muted (i.e., $\varepsilon_{w_{F_j},\mathcal{I}} = 0$). For these firms, the formal employment margin becomes more responsive.

In terms of formal employment, the corresponding expression is equal to:

$$\varepsilon_{F_j,\mathcal{I}} = \Omega_j s_{I_j}(\epsilon - \rho)\beta_F. \tag{11}$$

The implications for formal employment in terms of the substitution parameter (i.e., $\rho > \epsilon$) hold similarly as for formal wages, though the response is now adjusted by β_F (indicating that formal employment is more responsive than formal wages to informal supply shocks). Hence, in firms without informal workers, formal employment will not be adjusted as a response to the informal

[42]For simplicity, in this model I do not distinguish if the produced good is tradable or non-tradable, only that the firm produces a good.

[43]GEIH survey data shows that around 90% of Venezuelan immigrants are employed in the informal sector.

[44]To show that $\Omega_j > 0$, note that this is simplified as $\Omega_j = (1 + (1+\eta-\rho)\beta_I)(1 + (1-\rho)\beta_F) - (\epsilon-\rho)(s_{I_j}\beta_I + s_{F_j}\beta_F + (1+\eta-\rho)\beta_I s_{F_j}\beta_F + (1-\rho)\beta_F s_{I_j}\beta_I)$ which is always positive.

supply shock. Yet, as the relative contribution of informal workers to production increases ($s_{I_j} \uparrow$) the adjustment on formal employment is going to be more negative ($\varepsilon_{F_j,\mathcal{I}} \downarrow$) when formal and informal workers are close substitutes.

For the rest of the firms' responses, in Appendix E I show that the elasticity of informal labor is always positive ($\varepsilon_{I_j,\mathcal{I}} > 0$) and the elasticity of informal wages is always negative ($\varepsilon_{w_{I_j},\mathcal{I}} < 0$) after an aggregate informal supply shock. Independent if informal and formal workers are close substitutes or not.

To summarize, the proposed model points to two main conclusions. The first one is that to have a negative response on formal wages and formal employment, there needs to exist a high degree of substitutability between formal and informal workers. The second one is the importance of the production structure to determine how responsive the firm is to an informal supply shock. In particular, as firms' weight on informal workers for production is higher, it will be *more responsive* in terms of formal wages and employment to the shock.

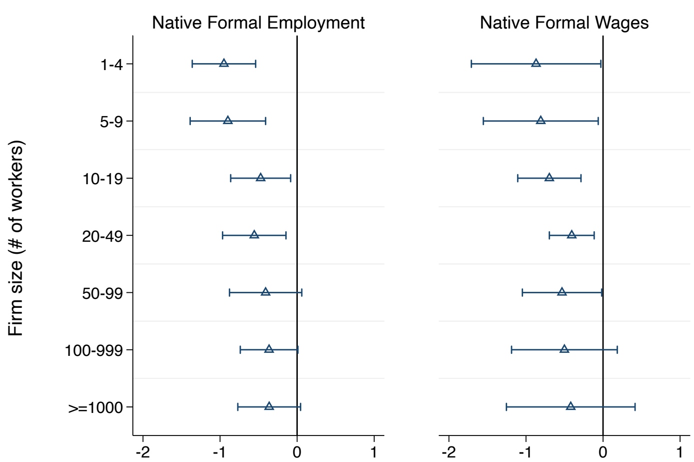
In the model, the firm's informal production share is inversely related to the firm's size. This is because for larger firms is marginally more expensive to hire an additional informal worker due to the convex cost of informal labor $\tau(I_J)$. For that reason, in the empirical results, I focus first on firms' size to show that the patterns predicted from the model are also observed in the data.

5.2 Worker-level Effects by Firm Characteristics

The previous model suggests that certain types of firms should react more adversely to an immigration shock, so I turn to the data to test these implications. In this context, using the firm component is also motivated by three stylized facts. First, Venezuelan immigrants are disproportionally employed in the smallest firms, second, small firms are more likely to pay the minimum wage for their formal workers and, third, small firms employ a higher share of informal workers (Delgado-Prieto, 2022). Hence, the impact of immigration on small firms is more salient. In this exercise, I divide workers by firm size categories in 2015 (the year before the immigration shock) and show worker-level employment and wage coefficients for 2018 (the year of the immigration shock from the census).

Figure 8 shows that native workers in firms with less than 50 workers in the pre-shock period suffer the most negative effect on the probability of being a formal worker, while workers in bigger firms are less affected. In line with the model's predictions, small firms tend to rely more on informal workers for production, and the cost of being caught by authorities is lower in these firms compared to the largest ones. Thus, when formal and informal workers are close substitutes, it is profitable for the firms to substitute them. There is suggestive evidence from Delgado-Prieto (2022), using survey data of informal and formal workers, that the share of informal workers increases more in smaller firms after the arrival of migrants, indicating a change in the composition of the firm's workforce. Regarding wages, native workers in firms with less than ten workers have the most negative coefficient, yet all workers with less than 100 have a significant negative effect. A similar pattern is predicted from the model, where smaller firms are more wage responsive to an immigration shock. Last, these results are useful to transparently show that trade shocks from the Venezuelan crisis are less of a concern in this setup, as the main effects are observed in the small firms that are directly affected by migration and presumably much less by trade.

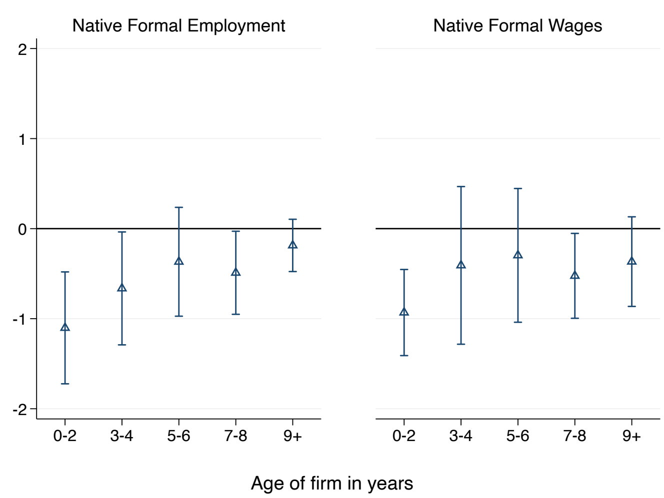
Figure 8: **Estimates by firm size, 2015–2018**



Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Workers are observed in August of each year. The coefficients for employment (in percentage points) and for wages (in percent) are already multiplied by 100. Source: PILA, 2013–2019.

Next, I quantify the immigration effects by exploiting other relevant firm characteristics. In this case, I exploit the years the firm appears in the administrative records up to the pre-shock period, that is, a proxy of the firm's age. Figure 9 shows results for native workers according to the age of their firm in 2015. For employment, workers in younger firms present a more negative impact than workers in older firms, while the pattern is not equally clear for wages. Still, workers in the youngest firms present the most negative coefficient on wages. The positive correlation between a firm's size and age helps explain previous negative findings, as smaller firms tend to be younger. To complement, Fort et al. (2013) show the different responses of young and old firms depending on their size during the business cycle, so I combine these characteristics to measure how immigration effects vary. Appendix Table B.2 shows that native workers in the youngest firms present a significant negative effect on employment and wages, independent if their firm is small or large, but the coefficient for wages is more negative in younger firms. On the other hand, native workers in older firms present a significant negative effect on employment and wages only in the smallest firms.

Figure 9: **Estimates by age of firm, 2015–2018**



Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to native employees between 25 and 55 years old. The firm's age is the number of years the firm appears discontinuously in PILA. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019.

5.3 AKM Decomposition

With access to the universe of workers and firms that contribute to the social security system in Colombia, it is possible to construct a measure of the overall quality of workers and firms. I estimate the standard AKM model proposed in Abowd et al. (1999) that decomposes the contribution of firm-specific and worker-specific constant characteristics to log formal wages (lnw_{it}). The AKM model is expressed as:

$$lnw_{it} = \alpha_i + \psi_{j(i,t)} + X'_{it}\beta + \epsilon_{it}. \tag{12}$$

Here, α_i captures the unobserved worker effect, ψ_j captures the unobserved firm effect, and $j(i,t)$ refers to the firm j where worker i is working in t. X_{it} is a vector of controls that are age squared and its cubic after being normalized and year FEs. Finally, ϵ_{it} is the error term.

To rule out possible endogenous movement of workers due to the immigration shock, I estimate the previous model from 2010 to 2015 for August ($T = 6$). It is required to identify the firm FEs that workers move across firms in the period of analysis and, for consistency, it is required the standard mean independence assumption $E[\epsilon_{it}|\alpha_i, \psi_j(i,t), X_{it}] = 0$. For my analysis, I estimate the vector of firm FEs $\hat{\psi}_1, ..., \hat{\psi}_J$ and worker FEs $\hat{\alpha}_1, ..., \hat{\alpha}_N$. AKM models present issues when estimating firm FEs in firms with limited mobility of workers across firms, especially in smaller firms (Card et al., 2018; Bonhomme et al., 2020). Several solutions have been proposed to address this limitation, one of them would be to eliminate all the small firms from the estimation.[45] Since the majority of migrants are working in small firms, I prefer to restrict the sample to the largest set of firms connected by the mobility of workers to reduce the concern of limited mobility bias.[46]

To begin with, Appendix Table D.2 shows the decomposition of the variance of wages $Var(lnw_{it})$ in the formal sector of Colombia using the leave-out method proposed by Kline et al. (2020). Worker effects explain 50.2% of the variance and firm effects explain 15.7%, in line with the related literature cited in Card et al. (2018). Furthermore, there is a positive sorting of high-wage workers into high-wage firms, which explains an additional 21.6% of the variance.

Then, using the estimated $\hat{\psi}_j$, I divide workers by seven quantiles of firm FEs or firm-specific pay premiums, which I now refer to as lowest- or highest-paying firms, to compute the impact of immigration. Figure G.1 shows that workers at low-paying firms in 2015 suffered the most negative employment losses while having insignificant wage changes. In contrast to workers in

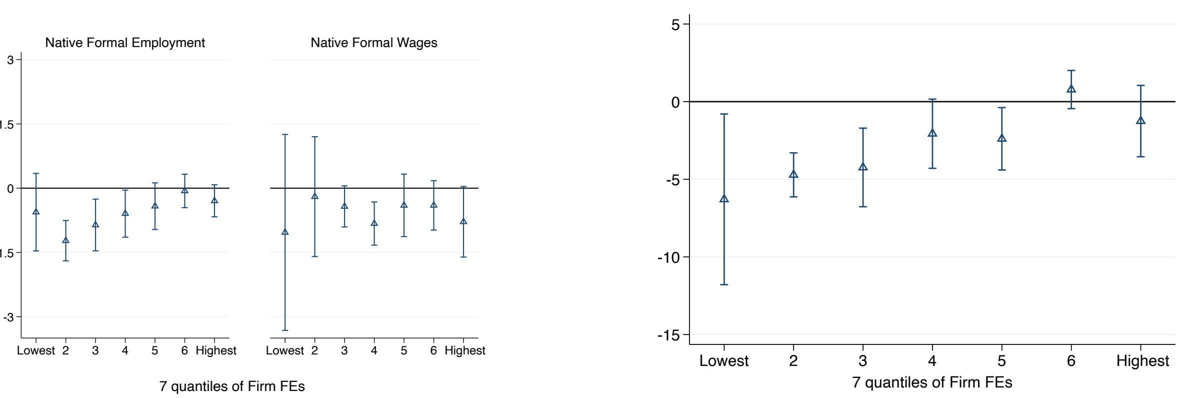
[45] Another solution is to aggregate small firms according to their observable characteristics, but as I observe industries with measurement error, the aggregation could include high-productivity sectors with low-productivity ones, misleading the estimates.

[46] The leave-out estimation of variance components in Kline et al. (2020) is a different solution to this problem. However, this method yields the corrected moments of interest (i.e., the variance of firm and workers FEs with their corresponding covariance) but does not estimate the corrected vector of $\hat{\psi}_j$ used in this paper.

middle-paying firms in 2015, where the wage response is as negative as the employment response. A possible explanation for this result is that the share of firms in the low-pay sector gets bigger because immigrants work mainly in these firms, and as a response, firms in the high-pay sector extract higher rents from workers and hence reduce their wages, as shown theoretically in Amior and Stuhler (2022). Last, to define which adjustment, if wages or employment, decreases more overall earnings, I exploit the cumulative earnings outcome. Figure 10b shows that workers in the lowest-paying firms are much more affected in terms of earnings than workers in high-paying firms.

Figure 10: **Estimates by quantiles of firm FEs, 2015–2018**

(a) Employment and wages (b) Cumulative earnings



Note: In (a), dependent variables are employment relative to the pre-shock period and wages relative to the base period. In (b) dependent variable is $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,2015}}$. The sample is restricted to natives between 25 and 55 years old. Firm FEs are computed in the first step using the standard AKM framework, with age squared and its cubic as time-varying controls, for the period 2010-2015. Controls used in the second step are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019.

Appendix Figure B.3 shows a similar exercise but dividing by seven quantiles of worker FEs: $\hat{\alpha}_i$. The wage and employment estimates hold similarly as before. Higher-quality workers present more negative point estimates for wages and the least negative ones for employment, in contrast to lower-quality workers, where the wage effect is close to zero, and the employment effects are more negative.It is important how wages and employment mirror each other and should be studied together as emphasized in Dustmann et al. (2017).

5.4 Heterogeneity by Worker and Firm Characteristics

As shown previously, certain workers' and firms' characteristics determine a differential impact of immigration on wages and employment. To illustrate the groups most affected in a more standard way, I restrict the sample to the intersection between subgroups where previous findings indicate a more negative coefficient. In the next section, I present a more systematic analysis of heterogeneity using a machine learning algorithm. First, Table 5 shows that for minimum wage earners in 2015, immigration reduced the probability of employment in the formal sector by 1.5 pp. For the medium age group, the impact is less negative (–1.2 pp), while for self-employed workers, the impact is more negative (–2.2 pp). When combining these three characteristics, there are 565,594 workers in the sample, for whom the effect of Venezuelan immigration on the probability of being a formal worker is -2.6 pp, a larger displacement effect.

Table 5: **Most affected native workers in terms of employment, 2015–2018**

	(1)	(2)	(3)	(4)	(5)
Prob. of Employment	-0.841***	-1.453***	-1.188***	-2.194***	-2.647***
	(0.192)	(0.231)	(0.227)	(0.327)	(0.388)
Sample restriction					
Minimum wage earners	✗	✓	✗	✗	✓
Medium age (35 years or more)	✗	✗	✓	✗	✓
Self-employed	✗	✗	✗	✓	✓
N	6,706,035	2,205,814	3,915,188	1,103,384	565,594
Clusters	109	109	109	109	109

Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The outcome variable is $e_{i,2018} - \sum_{t=2013}^{2015} e_{it}/3$ where e_{it} is the indicator of being employed in the formal sector. To understand how large the coefficients are, the size of the formal sector in urban areas, relative to overall employment, was 55.2% in 2015. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2013–2019.

Next, I use the same criteria as in Table 6 to divide the sample by the subgroups with the highest negative coefficient, but for native wages. First, I find that for workers earning more than the minimum wage in 2015, migration reduced average wages by 0.7%. For workers in the smallest firms in 2015, the impact is more negative (–0.8%), while for workers in middle-paying firms in 2015, I find an estimate of -0.8%. When combining these characteristics, there are 53,279 workers in the sample, for whom the effect on wages in 2018 is –1.9% for a one pp increase in the immigration shock.

Table 6: **Most affected native workers in terms of wages, 2015–2018**

	(1)	(2)	(3)	(4)	(5)
Wages	-0.600*	-0.711*	-0.827**	-0.804**	-1.908**
	(0.239)	(0.315)	(0.320)	(0.260)	(0.477)
Sample restriction					
Above minimum wage	✗	✓	✗	✗	✓
Small firm (1 and 19 workers)	✗	✗	✓	✗	✓
Middle-paying firm (quantile 4)	✗	✗	✗	✓	✓
N	4,090,973	2,639,040	643,346	195,647	30,772
Clusters	109	109	109	109	109

Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The outcome variable is $\frac{w_{i,2018} - w_{i,2015}}{w_{i,2015}}$ where w_{it} are wages in the formal sector. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2015–2018.

5.5 Sorting

Next, I study the reallocation effects of the immigration shock analyzing changes in the sorting patterns of high- and low-paying workers into high- and low-paying firms.[47] In this exercise, the outcome is constructed using the values of $\hat{\psi}_j$ from equation (12) and exploiting the movements of workers between firms in the post-treatment period. More concretely, the outcome is the change in the AKM firm FEs in 2018 relative to 2015: $\hat{\psi}_{i,\{j=2018\}} - \hat{\psi}_{i,\{j=2015\}}$. If the worker remains in the same firm during that period, the difference is zero.[48] Results are shown by seven quantiles of worker FEs to determine if low- or high-wage workers are sorting more into low- or high-paying firms after the immigration event. A positive coefficient means a positive sorting effect from immigration. Figure 11a plots the estimates for these categories, and none of them present significant results.[49] There is no differential sorting due to immigration.[50] Thus, to explain the negative wage coefficient of workers in high-paying firms, there must be lower wage growth within these firms. In a related

[47]For France, Orefice and Peri (2020) study the changes in worker-firm sorting after immigrants arrive, they find that high-paying workers are moving more into high-paying firms.

[48]Since the FEs are constructed for the pre-policy period, all workers that belong to firms created after 2015 are not considered in the analysis. Last, the estimated firm FEs are transformed into positive values to construct the outcome.

[49]Compared to Germany, the introduction of a nationwide minimum wage led to the reallocation of low-wage workers into higher-paying firms (Dustmann et al., 2022).

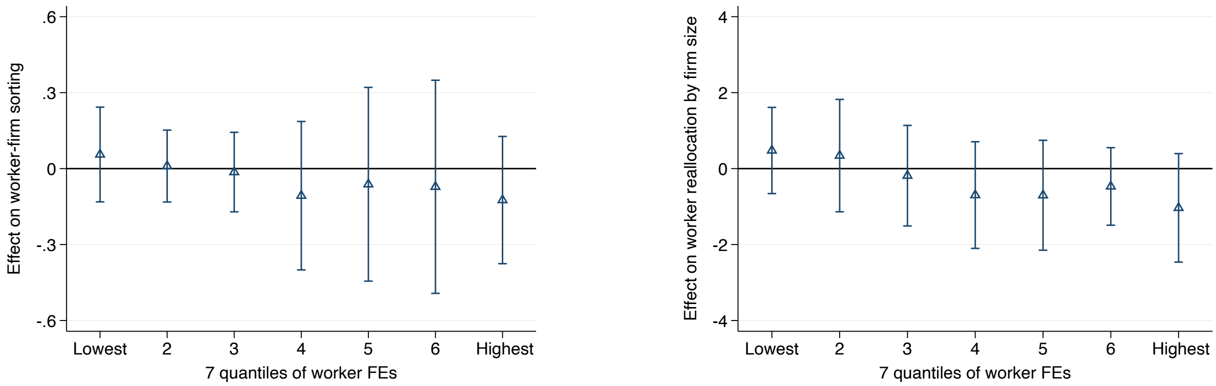
[50]This is partly attributed to the macroeconomic conditions of the labor market in Colombia during the period studied, as unemployment slightly increased.

exercise, I also measure if workers are moving to larger or smaller firms after the immigration shock, and again, there does not seem to be reallocation on this margin (see Figure 11b).

Figure 11: **Reallocation estimates by quantiles of worker FEs, 2015–2018**

(a) By firm pay-premiums

(b) By firm size



Note: Dependent variable in (a) is the change in $\hat{\psi}_{i,\{j=2018\}} - \hat{\psi}_{i,\{j=2015\}}$ and in (b) is the change in the categories of firm size in 2018 relative to 2015, both measured in the base period. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Workers are observed in August of each year. Source: PILA, 2013–2019.

5.6 Hiring Patterns of Formal Firms

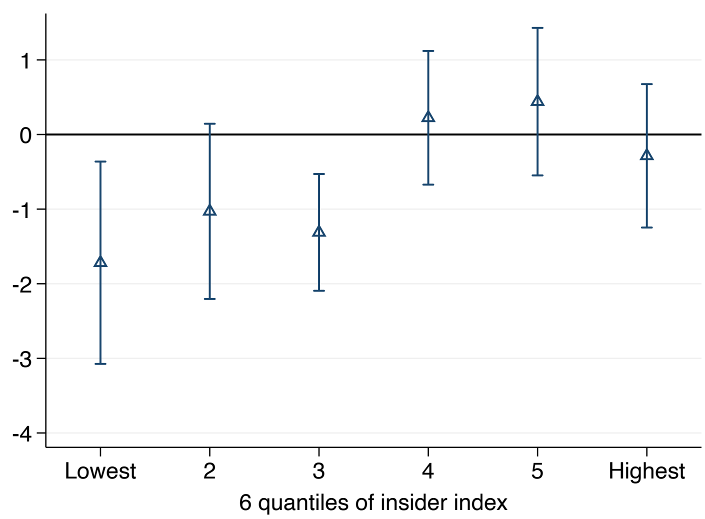
In the absence of informal worker-level data in the PILA records, it is possible to construct a measure of connectedness with the informal sector for formal firms apart from the standard firm size variable. To build this proxy, I develop an insider index similar to the poaching index constructed in Bagger and Lentz (2019). The intuition of the index is that firms are divided by the share of hires that come from *outside* the formal sector (most probably from the informal sector) and from *inside* the formal sector.[51] In some way, it is a measure of revealed preferences of workers, as firms that hire more from the formal sector are more desirable, while the firms that hire less act as "gatekeepers"' for workers that have not been employed in the formal sector before. The insider index is constructed for every firm j before and after immigrants arrive,

$$\pi_{j,t} = \frac{N_{j,t}^{In}}{N_{j,t}^{In} + N_{j,t}^{Out}}, \tag{13}$$

[51] Workers inside the formal sector may be employed in other firms at the time of the hire or may be unemployed but have worked previously in the formal sector.

where $N^{In}_{j,t}$ is the number of firm j in year t hires that have been employed before in the formal sector, and $N^{Out}_{j,t}$ is the number of firm hires that come outside the formal sector.[52] Next, I take differences in the insider index between 2018 and 2015 (i.e., $\pi_{j,2018} - \pi_{j,2015}$) at the worker level according to the firm the worker was employed in 2015. With this in mind, Figure 12 shows results for this outcome by six quantiles of the insider index in the pre-shock period. Interestingly, formal firms that hire more from outside the formal sector are having a negative reduction on their insider index after immigrants arrive, indicating that new hires in these firms are coming more from outside the formal sector (a 1 pp increase in the migration rate reduces the insider index of the lowest type of firms by around 1.7 pp). On the opposite, for firms that have a higher share of hires within the formal sector, the insider index does not change much. This measure is an important way of showing that given firms are opting out from the formal sector for new hires, especially the ones that are supposedly more connected to the informal sector, according to the insider index.

Figure 12: **Estimates by quantiles of the insider index, 2015–2018**



Note: Dependent variable is the change in the insider index for workers employed in firm j in the base period between 2015-2018. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories in the base period. Workers are observed in August of each year. Source: PILA, 2013–2019.

5.7 Exit and Entry of Formal Firms

In a related exercise, I test how likely it is that firms disappear entirely from the formal sector after immigrants arrive. Since workers are being displaced from the formal sector, it is unclear

[52] I can record the hires of firms since 2007 and can build the measure up until 2018 for February and August in each year. If the firm did not make any hiring in the year, the index takes a missing value.

if the displacement of certain workers, intensive margin, also happens for all the workers in the firm, extensive margin. Table 7 shows evidence that the growth in the number of firms in the formal sector is smaller in places that receive more migrants, yet the coefficient is insignificant. If I decompose the growth in the exit and entry margin of firms, there is a significantly higher firm exit. A 1 pp increase in the migration shock increases the firm exit rate by 1.2%. On the other hand, the firm entry rate is close to zero, indicating that the growth in new formal firms in places that receive more migrants is not different from places that receive fewer migrants.

Table 7: **Decomposition of Firm Growth, 2015–2018**

	(1) Total Firms	(2) Firm Exit	(3) Firm Entry
$\Delta M_{l,2018}$	-1.127	1.190*	0.063
	(0.750)	(0.582)	(0.935)
N	109	109	109

Standard errors are in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Regressions are estimated at the regional level for 109 FUAs weighted by their formal employment in 2015. The outcome variable in (1) is the growth in the number of firms, while in (2) and (3), I decompose the growth in terms of the exit and entry of firms, respectively. The sample is restricted to firms with at least one native employee. Firms are observed in August of each year. Source: PILA, 2015–2018.

6 Machine-Learning Approach

In this section of the article, I develop a machine learning algorithm to identify the subgroups most affected by immigration and to determine a proxy for the role of firms in the labor market effects of immigration. In previous sections, wage and employment effects are shown for arbitrarily chosen subgroups of the population according to given characteristics, yet to determine exactly which variable explains most of the heterogeneity of immigration effects, I turn to a data-driven approach proposed by Athey and Imbens (2016) and generalized in Athey et al. (2019). Recently, it was implemented by Gulyas et al. (2019) and Yakymovych et al. (2022). This framework identifies the subgroups that experience the greatest wage and employment losses by a recursive partitioning method while allowing for nonlinear effects and high-order interactions between firm and worker variables. The generalized random forest (GRF) method in Athey et al. (2019) builds causal forests, in the spirit of random forests (Breiman, 2001) but splits the data according to a criterion on treatment effect heterogeneity.[53] The benchmark specification that the algorithm uses is the following:

$$\Delta Y_{i,l,2018} = \tau(x_i)\Delta\hat{M}_{l,2018} + \Delta\epsilon_{i,l,2018} \tag{14}$$

where x_i are the values of the variables in X_i and $\tau(x_i)$ is the treatment effect. Moreover, ΔY is the outcome of interest, the difference of individual wages or employment in 2018 relative to the pre-shock period, that matches the census rate, and I use the predicted immigration rate $\hat{M}_{l,2018}$ after regressing the observed one with the two instruments. This is done because the algorithm does not allow more than one instrument. Vectors of worker and firm characteristics, including the ones constructed from the AKM model, are the partitioning variables of X_f. All these features or variables correspond to characteristics in 2015 (before the immigration shock), and they are age, sex, job tenure, wages, firm FEs, worker FEs, and firm size.[54] Self-employed workers are omitted in this section due to their incomparable information in most of the firm characteristics to employees.

The procedure in Athey and Imbens (2016) and Athey et al. (2019) to build causal trees consists of several steps that are adapted to this setup. Broadly, the algorithm proceeds as follows:

1. Start with 50% of the full sample P.[55] The remaining out-of-bag (OOB) sample will be used for prediction after the algorithm is trained.
2. Take a random subsample, without replacement, of P and choose a variable randomly from X_f and a value, from all possible values, for this selected variable.
3. For every possible value of one variable in X_f, the data is split into two partitions (say P_l and P_r) to run separate regressions of form (14) to estimate treatment effects for each partition. Choose the variable with its cutoff value that maximizes the difference in treatment effects using this formula:
$$(\tau_l - \tau_r)^2. \tag{15}$$
[56]
4. Observations with a value below or equal to the cutoff value are placed into a new left node, and observations with a value above are placed into a new right node of the decision tree.

[53] I use the `grf` package in `R` to estimate the causal forests.

[54] The procedure sample varies depending on the features selected but starts with the same sample. For instance, to construct worker effects, the individual needs to be observed more than once in the sample, so in this case, the sample is smaller.

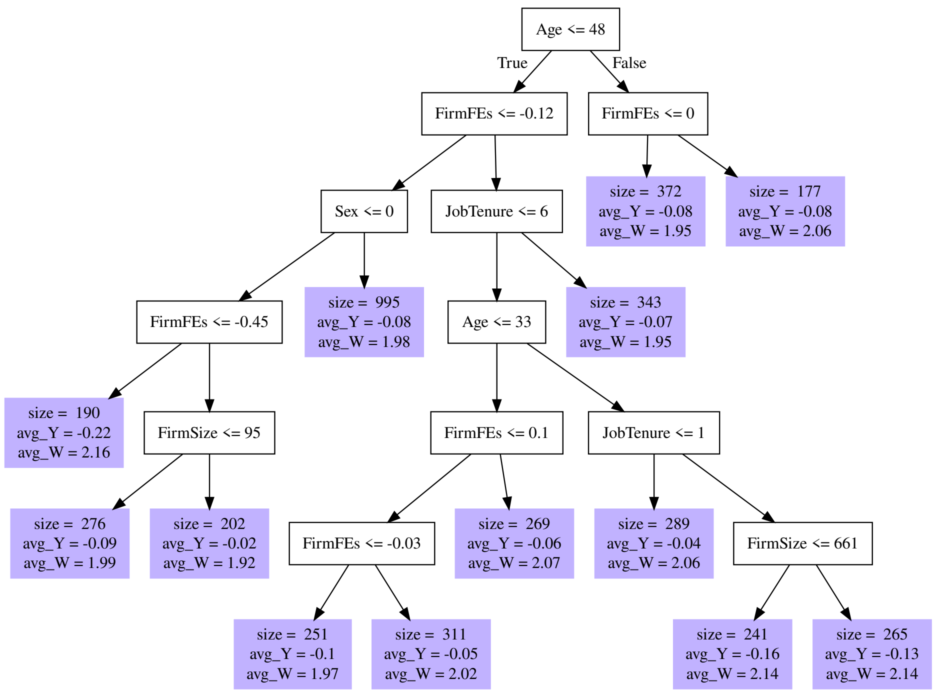
[55] The 50% threshold is selected due to computational burden. This subsample is further cut by 50% to do sub-sample splitting to create similarity matrices.

[56] There are penalties in the algorithm for the imbalance of the splits. For instance, the squared difference criterion can include an additional term $\frac{n_l n_r}{N^2}$ to adjust for more balanced splits (n_l and n_r refer to the sample size of each partition, and total subsample refers to N).

5. Recursively forms the resulting nodes with this algorithm until the nodes reach a minimum node size, the difference in sample size between the two partitions is large, or when the split would only yield a difference in treatment effects relatively small.

As an illustration of a decision tree in the causal forest algorithm, I use a 1% random sample of the main data. Figure 13 shows how observations of certain characteristics are placed to the right and the left of the tree. For the main algorithm, the causal forest is estimated using 2,000 decision trees with a minimum node size of 300, while clustering observations in FUAs.[57] Having many trees with a minimum node size reduces overfitting concerns.

Figure 13: **Illustration of decision tree**



Note: Dependent variable Y is employment changes in 2018 relative to the pre-shock period, and the predicted migration shock in 2018 is W. This decision tree uses a 1% random sample of the data.

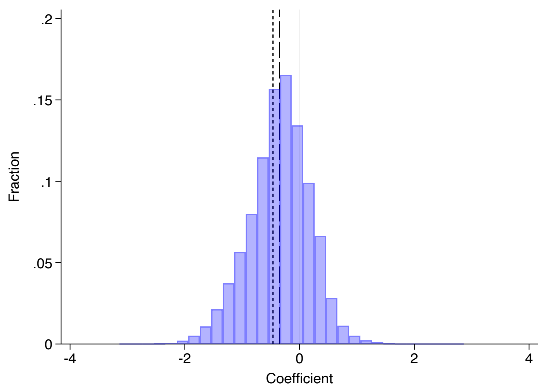
The first output of this procedure is shown in Figures 14a and 14b. According to the trained causal forest, these histograms plot the predicted individual treatment effects for both outcomes, wages, and employment. These treatment effects come from the OOB sample not used in the algorithm. More concretely, each OOB observation is classified in this forest into a final node of each tree. Then, for all trained trees, it counts the times these observations fall in the same

[57]The tunable parameters from the algorithm are set to default values, including the honest splitting, while the selected minimum node size is fairly small for precision. In a further cross-validation exercise, results hold when substantially increasing the minimum node size.

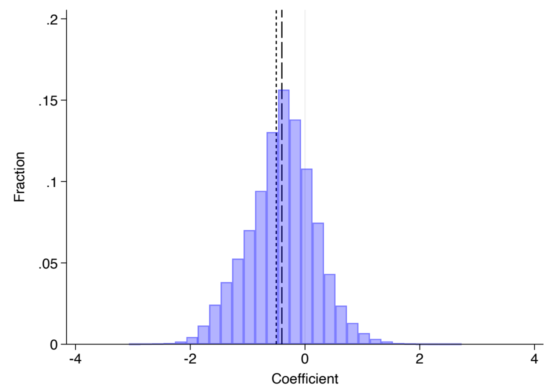
terminal node as the training sample to calculate the similarity weights. Using these weights, get the weighted mean of τ across trees for every observation to calculate the individual treatment effect $\tau(x_i)$. In the histograms, the average individual treatment effect is the long dashed line, and the average treatment effect from a standard regression is the short dashed line. For both outcomes, the average coefficient from the causal forest is similar, reflecting the accuracy of the predictions.

Figure 14: **Histogram of treatment effects for formal employment and formal wages in the causal forest, 2015–2018**

(a) Employment



(b) Wages



Note: The short dashed line refers to the coefficient from the benchmark specification, and the long dashed line refers to the average predicted treatment effects that are estimated with the trained causal forest using the OOB sample. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. I use clusters at the FUA level for the causal forest. The causal forest uses 50% of the main sample due to computational burden. The minimum node size is 300.

Next, I exploit the construction of treatment effects from the algorithm to describe which subgroups are most affected by immigration. In this exercise, native workers are divided into quintiles of treatment effects of employment and wages (quintile 1 refers to the most negative coefficient and quintile 5 to the most positive one). Tables 8a and 8b show worker and firm characteristics in the pre-shock period. First, native workers with the most negative employment effects are the oldest, with the lowest tenure, and earning the lowest initial wages. Besides, these workers are employed in the smallest firms with the lowest pay premiums. Conversely, workers that suffer the most negative wage effect are relatively younger, with few years of tenure, and earn the highest initial wages. In terms of firm characteristics, these workers are employed in the smallest firms, and in terms of pay premiums, they are in the middle-high part.[58] From a policy perspective,

[58]In Appendix Figures C.3a and C.3a, I test if the quintiles of treatment effects from the causal forest yield the

the distribution of individual treatment effects is useful for targeted measures that aim to decrease the losses from immigration in the most affected subgroups.

Table 8: **Descriptive statistics for native workers by quintiles of treatment effects**

(a) Formal employment

	Q1	Q2	Q3	Q4	Q5
Male (%)	0.7	0.6	0.5	0.5	0.5
Age of worker	42.8	40.3	38.5	35.1	31.1
Job tenure (1-9 years)	2.3	3.6	4.4	4.1	2.8
Monthly wages (USD)	324.8	462.6	521.8	478.4	336.2
Median firm size	79	105	276	510	1109
Quantiles of firm FEs (1-7)	3.8	5.3	6.0	6.3	6.5

(b) Formal wages

	Q1	Q2	Q3	Q4	Q5
Male (%)	0.6	0.6	0.6	0.6	0.5
Age of worker	36.6	38.5	38.8	38.1	37.5
Job tenure (1-9 years)	3.2	3.9	4.0	3.8	3.5
Monthly wages (USD)	559.5	466.2	419.3	379.0	393.7
Median firm size	86	189	242	309	892
Quantiles of firm FEs (1-7)	5.7	5.8	5.6	5.5	5.5

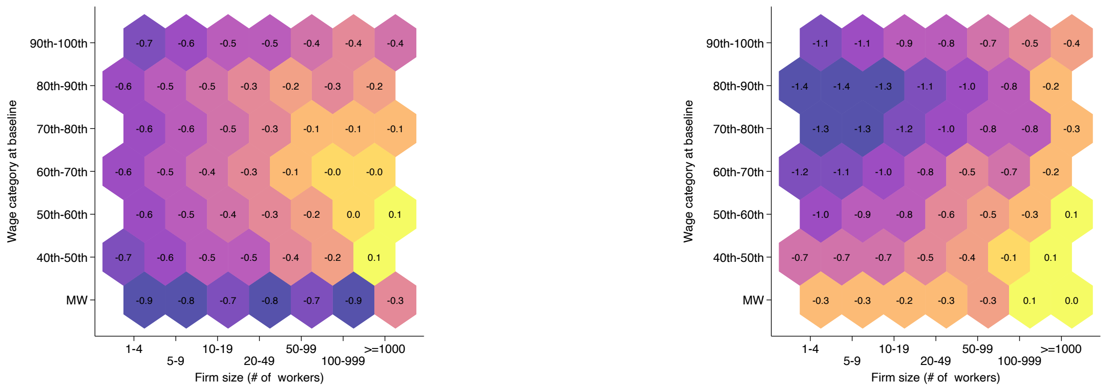
Note: These tables report the average or median statistics for quintiles of treatment effects (Q1 is the most affected and Q5 is the least affected), in terms of employment and wages, according to the predictions of the trained causal forest using the OOB sample. The wages are transformed from Colombian pesos to USD using 2020 exchange rates from the World Bank. Source: PILA, August 2015.

Following up, and to have a better illustration of the subgroups most affected by immigration, I construct heat plots. Figures 15a and 15b show average treatment effects by individual wages at baseline for different firm sizes. Interestingly, most negative employment effects are concentrated on the intersection of minimum wage earners employed in small and medium firms. Opposite from the most negative wage effects, which are concentrated in the upper part of the wage distribution, but again in small firms (wage effects are smoothly disappearing as the firm grows).

same order when using the main empirical specification. Importantly, the estimates follow the same order as the quintiles for wages and employment.

Figure 15: **Heat plot of average treatment effects by wage category at baseline and firm size, 2015–2018**

(a) Employment (b) Wages

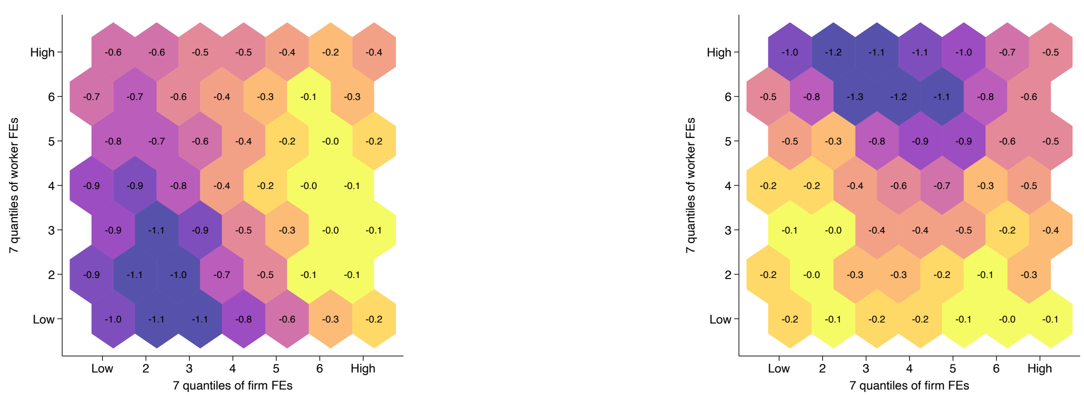


Note: Each hexagon is the average treatment effect for that subgroup according to the trained causal forest using the OOB sample. The sample is restricted to natives between 25 and 55 years old. I use clusters at the FUA level for the causal forest. The causal forest uses 50% of the main sample due to computational burden.

Next, Figures 16a and 16b show average treatment effects by quantiles of firm FEs intersected with quantiles of worker FEs. Interestingly, most negative employment effects are concentrated on the lowest-quality workers in the lowest-paying firms. Opposite from the most negative wage effects, which tend to be concentrated in higher quality workers in middle-paying firms.

Figure 16: **Heat plot of average treatment effects by quantiles of workers and firm FEs, 2015–2018**

(a) Employment (b) Wages

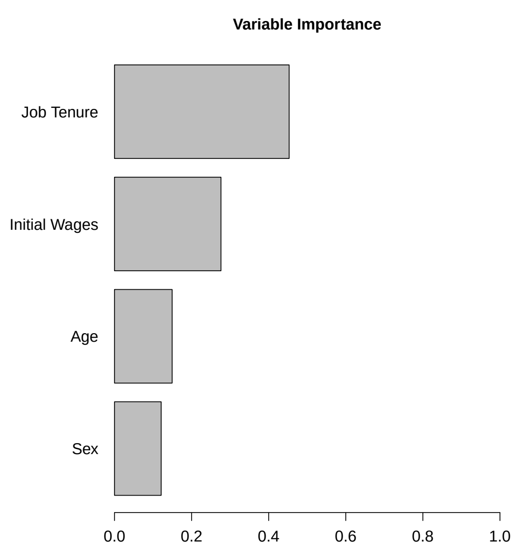


Note: Each hexagon is the average treatment effect for that subgroup according to the trained causal forest using the OOB sample. The sample is restricted to natives between 25 and 55 years old. I use clusters at the FUA level for the causal forest. The causal forest uses 50% of the main sample due to computational burden.

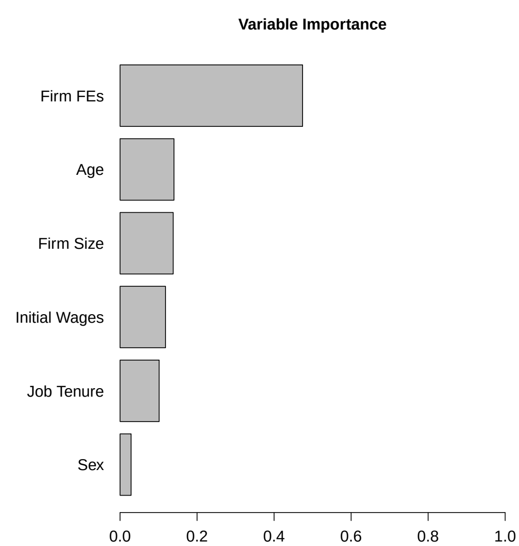
A complementary way of summarizing these findings is with the variable importance measure. In this case, the variables that appear more frequently as splits in the forest are categorized as more important to explain treatment effect heterogeneity. This naive measure yields a ranking that serves as a proxy to classify the sources of heterogeneity. To start, I perform the algorithm excluding and including firms' variables to show how different is the importance measure.[59] Hence, when excluding firms' variables, I find that job tenure, followed by initial wages and age, are more important to determine the heterogeneity on employment impacts of migration (see Figure 17a). However, when including firms' variables, the most important variable becomes firm-specific pay premiums or firm FEs, followed by age and firm size. In contrast, the least important one in both cases is sex (see Figure 17b). Thus, the relevance of firms for the heterogeneity of immigration effects on natives is notable. Like the findings of Arellano-Bover and San (2020), that shows the important role of firms in the assimilation of immigrants in the labor market.

Figure 17: **Variable importance for formal employment in the causal forest, 2015–2018**

(a) Without Firm Variables



(b) With Firm Variables



Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. I use clusters of FUA for the causal forest estimation. The minimum node size is 300.

Following up, I use the individual wage growth between 2018 and 2015 to perform the same

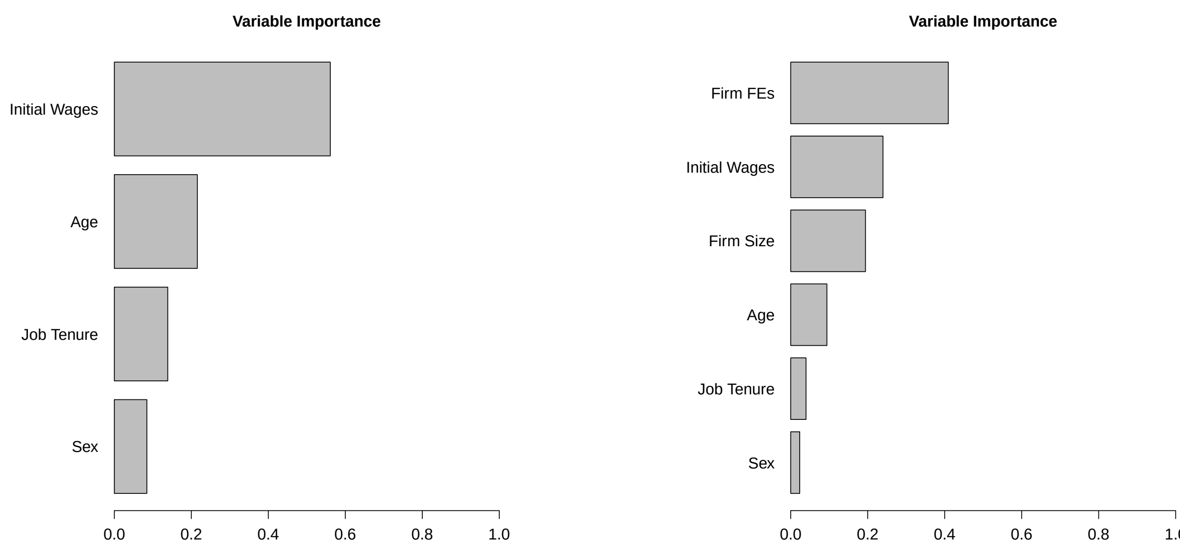
[59] For employment, I use the individual change in employment between 2018 and the average pre-shock period employment as the outcome.

exercise. Without firms' variables, the most important variables are initial wages followed by age and job tenure (see Figure 18a). However, when including firms' variables in the causal forest, firm-specific pay premiums followed by initial wages and firm size are the most important (see Figure 18b). Note that the variables of firm size and pay premiums are positively correlated, but the correlation is not so strong (0.19). Conversely, the variables that explain the least are job tenure and sex. To summarize, the most important variable to explain wage and employment changes relates to the firm-specific pay premiums or firm FEs more than any worker characteristics. In the causal forest of wages, firm effects appear in 37% of all splits; for employment, firm effects appear in 30% of the splits.

Figure 18: **Variable importance for formal wages in the causal forest, 2015–2018**

(a) Without Firm Variables

(b) With Firm Variables



Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. I use clusters of FUA for the causal forest estimation. The minimum node size is 300.

Last, as initial wages are a function of the unobserved firm and worker FEs, the next step is to include in the algorithm the constructed worker FEs $\hat{\alpha}_i$ instead of initial wages. This reduces the sample as every worker must be observed more than once. After adding worker FEs, again, the firm FEs are the most important variable to explain the heterogeneity of treatment effects for employment and wages (see Appendix Figures B.4a and B.4b). In conclusion, firms' role in the impact of immigration is still very relevant even after conditioning with the quality of workers.

7 Robustness Checks

The exclusion restriction of the distance instrument defines a unique causal path to the outcomes that are through the immigration shock. In that sense, border areas might fail this restriction as they are more prone to be affected by time-varying shocks arising from the Venezuelan crisis. Therefore, I remove all the border areas from the estimation sample to find similar point estimates but not significant for wages (see Appendix Table C.1, row 2). Next, another concern is the relevance of Bogotá as the capital of Colombia (the proportion of observations from the capital is 32.7% of the whole sample); hence I also remove it from the estimation sample and find that coefficients are less negative, especially for wages, but both are significant (see Appendix Table C.1, row 3).

Next, further controls are added to the estimation to better compare workers across local labor markets. The additional controls are seven groups of wage categories, according to the local wage distribution in 2015, and job tenure. Reassuringly, results are similar for wages but much less negative for employment, mainly because self-employed workers are omitted from the analysis as there is no comparable measure of job tenure. The next robustness test relates to adjusting nominal wages to real terms using the national CPI. In this case, the results of wages are slightly less negative. Last, to omit outliers driving the wage results, I top code wages after the 99% percentile of the wage distribution to find that estimates are unaltered.

Third, there are robustness checks for the machine learning algorithm. The first one is that firm pay-premiums might be correlated with the type of industry the firm belongs to, reflecting that some industries generally have higher or lower wage premia (Card et al., 2022). For that reason, I include in the algorithm the industry of the firm, along with the firm FEs, to find that for wages and employment, the most important variable is still the firm FEs (see Appendix Figures C.5a and C.5a). The second one deals with the fact that the frequency of splits in the first nodes of the trees is weighted the same as the frequency of splits in the last nodes of the tree, where the sample size is much smaller. This critique is alleviated by using a decay exponent in the variable importance that puts more weight on the splits selected first.[60] After computing the variable importance, the order is fairly similar for wages and for employment firm size is now the second most important variable, preceded by firm FEs; interestingly, both variables capture the role of firms (see Appendix Figures C.4a and C.4a).

Next is that, in the causal forests, the number of possible values a variable takes might alter the variable importance weighted sum (Strobl et al., 2007). For instance, when variables have a small set of values, they might mechanically appear in fewer nodes further in the tree. For that reason, Appendix Figures C.6a and C.6b show that when transforming all the continuous variables into seven or six categories, as the ones in previous results, the order of importance is similar for employment but for wages changes slightly, with firm FEs being second. Still, even if the main results hold, the benefit of the algorithm comes from exploiting all the possible values a variable takes, not from arbitrarily aggregating into categories. Another critique is that the tree is built to maximize the squared difference in treatment effects without analyzing whether pre-trends are significant for all these subgroups, it just assumes strict exogeneity of the instrument. Probably when the treatment effects are higher, there could be differing pre-trends. However, as the algorithm is constructed, it does not allow correcting or checking for pre-trends in every possible subgroup. Finally, there is a recent statistical literature that proposes hypothesis testing of variable importance measures of random forests (see, for instance, Hapfelmeier et al. (2023)). The main idea is to perform sequential permutation tests to get the p-value of each variable used in the algorithm. The main issue is that it has not been developed for causal forests, and even if available, it is inefficient in this high-dimensional setup and computationally infeasible.

8 Conclusion

This is the first paper that exploits the labor supply shock of migrants from Venezuela equipped with data covering the universe of formal workers and firms in Colombia. This is an advantage in several dimensions. First, with administrative panel data, it is possible to follow workers over time and address compositional changes that arise in the standard regional-level analysis of immigration using survey data. Second, with the matched employee-employer dimension of the data, it is possible

[60] The decay exponent is -2, meaning that split frequencies in node k are weighted 1/2 compared to those in node $k-1$.

to exploit a large set of heterogeneous effects by the firm and worker characteristics that help to understand additional mechanisms after labor supply shocks in developing countries. Third, with the full count of formal firms and a machine learning method, it is possible to construct a proxy measure of firms' role in the impact of immigration on workers' outcomes.

Overall, the findings suggest that after migrants arrive, there is a negative impact on individual employment in the formal sector. However, this coefficient masks many heterogeneous responses. Specifically, minimum-wage workers are crowded out from the formal sector, while workers above in the wage distribution are not displaced, but instead, they have negative wage growth. Regarding firm characteristics, the negative effect on employment and wages is concentrated in small firms, which aligns with the predictions from the model of heterogenous firms that state that firms that use relatively more informal workers in production are more affected. Next, I find that workers in middle-paying firms present a negative wage adjustment, while workers in low-paying firms do not experience a reduction in wages but in employment, partly due to the existence of a relatively high minimum wage.

To uncover the mechanisms behind these estimates, I use causal forests to classify which variable is most important to explain the heterogeneity in treatment effects. Throughout this analysis, firm-specific pay premiums appear prominently as the most important variable to explain wage and employment losses, followed by firm size in most cases. Thus, using only workers' characteristics when analyzing the labor market impacts of immigration might lead to an incomplete view of the sources of adjustments to immigration. Indicating that after migrants arrive, the focus should not be, at least for Colombia, in *who* the worker is but more on which *type of firm* the worker is employed.

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Aksu, E., Erzan, R., and Kırdar, M. G. (2018). The impact of mass migration of Syrians on the Turkish labor market. Technical report, Working Paper.

Altonji, J. G. and Card, D. (1991). The effects of immigration on the labor market outcomes of less-skilled natives. In *Immigration, trade, and the labor market*, pages 201–234. University of Chicago Press.

Amior, M. and Stuhler, J. (2022). Immigration and monopsony: Evidence across the distribution of firms. Technical report, Working paper.

Arellano-Bover, J. and San, S. (2020). The role of firms in the assimilation of immigrants. *Available at SSRN 3594778*.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Autor, D. H., Dorn, D., Hanson, G. H., and Song, J. (2014). Trade adjustment: Worker-level evidence. *The Quarterly Journal of Economics*, 129(4):1799–1860.

Aydemir, A. and Borjas, G. J. (2011). Attenuation bias in measuring the wage impact of immigration. *Journal of Labor Economics*, 29(1):69–112.

Bagger, J. and Lentz, R. (2019). An empirical model of wage dispersion with sorting. *The Review of Economic Studies*, 86(1):153–190.

Bahar, D., Ibáñez, A. M., and Rozo, S. V. (2021). Give me your tired and your poor: Impact of a large-scale amnesty program for undocumented refugees. *Journal of Development Economics*, 151:102652.

Bonhomme, S., Holzheu, K., Lamadon, T., Manresa, E., Mogstad, M., and Setzler, B. (2020). How much should we trust estimates of firm effects and worker sorting? Technical report, National Bureau of Economic Research.

Borjas, G. J. (2006). Native internal migration and the labor market impact of immigration. *Journal of Human resources*, 41(2):221–258.

Borjas, G. J. and Edo, A. (2021). Gender, selection into employment, and the wage impact of immigration. Technical report, National Bureau of Economic Research.

Bratsberg, B. and Raaum, O. (2012). Immigration and wages: Evidence from construction. *The economic journal*, 122(565):1177–1205.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brinatti, A., Chen, M., Mahajan, P., Morales, N., and Shih, K. Y. (2023). The impact of immigration on firms and workers: Insights from the h-1b lottery. *Available at SSRN 4431106*.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19(1):22–64.

Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, 36(S1):S13–S70.

Card, D., Cardoso, A. R., and Kline, P. (2016). Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly journal of economics*, 131(2):633–686.

Card, D., Heining, J., and Kline, P. (2013). Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics*, 128(3):967–1015.

Card, D., Rothstein, J., and Yi, M. (2022). Industry wage differentials: A firm-based approach.

Unpublished draft, University of California, Berkeley.

Caruso, G., Canon, C. G., and Mueller, V. (2021). Spillover effects of the venezuelan crisis: migration impacts in colombia. *Oxford Economic Papers*, 73(2):771–795.

Ceritoglu, E., Yunculer, H. B. G., Torun, H., and Tumen, S. (2017). The impact of syrian refugees on natives? labor market outcomes in turkey: evidence from a quasi-experimental design. *IZA Journal of Labor Policy*, 6(1):1–28.

Corbi, R., Ferraz, T., and Narita, R. (2021). Internal migration and labor market adjustments in the presence of nonwage compensation.

DANE (2019). Gran Encuesta Integrada de Hogares (GEIH)-Módulo de Migración. https://microdatos.dane.gov.co/index.php/catalog/641/get_microdata.

Del Carpio, X. V. and Wagner, M. C. (2015). The impact of Syrian refugees on the Turkish labor market. *World Bank policy research working paper*, (7402).

Delgado-Prieto, L. (2022). Immigration, wages, and employment under informal labor markets.

Dostie, B., Li, J., Card, D., and Parent, D. (2021). Employer policies and the immigrant–native earnings gap. *Journal of Econometrics*.

Dustmann, C., Lindner, A., Schönberg, U., Umkehrer, M., and Vom Berge, P. (2022). Reallocation effects of the minimum wage. *The Quarterly Journal of Economics*, 137(1):267–328.

Dustmann, C., Otten, S., Schönberg, U., and Stuhler, J. (2023). The Effects of Immigration on Places and Individuals - Identification and Interpretation. Technical report.

Dustmann, C., Schönberg, U., and Stuhler, J. (2017). Labor supply shocks, native wages, and the adjustment of local employment. *The Quarterly Journal of Economics*, 132(1):435–483.

Foged, M. and Peri, G. (2016). Immigrants' effect on native workers: New analysis on longitudinal data. *American Economic Journal: Applied Economics*, 8(2):1–34.

Fort, T. C., Haltiwanger, J., Jarmin, R. S., and Miranda, J. (2013). How firms respond to business cycles: The role of firm age and firm size. *IMF Economic Review*, 61(3):520–559.

Gerard, F., Lagos, L., Severnini, E., and Card, D. (2021). Assortative matching or exclusionary hiring? the impact of employment and pay policies on racial wage differences in brazil. *American Economic Review*, 111(10):3418–3457.

Groeger, A., León-Ciliotta, G., and Stillman, S. (2022). Immigration, labor markets and discrimination.

Gulyas, A., Pytka, K., et al. (2019). Understanding the sources of earnings losses after job displacement: A machine-learning approach. Technical report, University of Bonn and University of Mannheim, Germany.

Hapfelmeier, A., Hornung, R., and Haller, B. (2023). Efficient permutation testing of variable importance measures by the example of random forests. *Computational Statistics & Data Analysis*, page 107689.

Kleemans, M. and Magruder, J. (2018). Labour market responses to immigration: Evidence from internal migration driven by weather shocks. *The Economic Journal*, 128(613):2032–2065.

Kline, P., Saggio, R., and Sølvsten, M. (2020). Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898.

Lebow, J. (2021). Immigration and occupational downgrading in colombia. *Unpublished Manuscript*.

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review*, 107(8):2278–2307.

Monras, J. (2020). Immigration and wage dynamics: Evidence from the mexican peso crisis. *Journal of Political Economy*, 128(8):3017–3089.

Morales-Zurita, L. F., Bonilla-Mejía, L., Hermida, D., Flórez, L. A., Bonilla-Mejía, L., Morales, L. F., Hermida-Giraldo, D., and Flórez, L. A. (2020). The labor market of immigrants and non-immigrants evidence from the venezuelan refugee crisis. *Borradores de Economía; No. 1119.*

Muñoz, M. (2021). Trading non-tradables: The implications of europe?s job posting policy.

Orefice, G. and Peri, G. (2020). Immigration and worker-firm matching. Technical report, National Bureau of Economic Research.

Sanchez-Serra, D. (2016). Functional urban areas in colombia.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1–21.

Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from brazil. *American Economic Review*, 108(8):2015–47.

UNHCR (2019). Venezuela situation. Technical report, Office of the United Nations High Commissioner for Refugees. Available on: https://www.unhcr.org/venezuela-emergency.html (last accessed 13th of June 2020).

Yagan, D. (2019). Employment hysteresis from the great recession. *Journal of Political Economy*, 127(5):2505–2558.

Yakymovych, Y., Nordström Skans, O., Vikström, J., Simon, L., and Athey, S. (2022). Worker attributes, aggregate conditions and theimpact of adverse labor market shocks.

Online Appendix

A First stage of the instruments

Table A.1: **First stage: The inflow of Venezuelan immigrants and the two instruments**

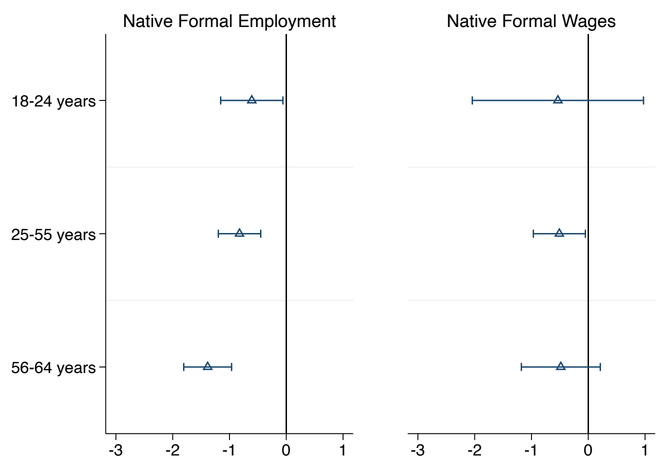
	(1)	(2)	(3)
	$\Delta M_{l,2018}$	$\Delta M_{l,2018}$	$\Delta M_{l,2018}$
Distance (/100)	-1.992***		-1.455***
	(0.272)		(0.350)
Distance (/100) squared	0.151***		0.107***
	(0.024)		(0.029)
Past settlements		0.703***	0.280*
		(0.160)	(0.130)
Constant	6.762***	1.040***	5.184***
	(0.715)	(0.149)	(1.000)
R^2	0.583	0.450	0.618
F	34.53	19.37	23.68
N	109	109	109

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table reports the coefficient of the first-stage of the share of employed migrants $\Delta M_{l,2018} * 100$ with distance and distance squared to the nearest crossing bridge and past settlements as explanatory variables.

B Additional Results

Figure B.1: **Estimates by extended age categories, 2015–2018**



Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to natives between 18 and 64 years old. Controls used are sex with a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and wages (in percent) are already multiplied by 100. Workers are observed in August of each year. Source: PILA, 2013–2019.

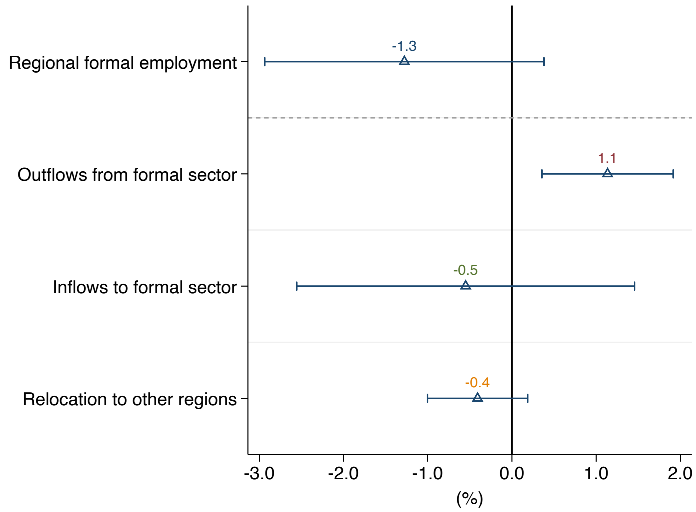
Table B.1: **Employment and wage estimates by age and job tenure, 2015–2018**

Worker's age	Below 35 years		Above 35 years	
Job tenure	0 to 4 years	5 to 9+ years	0 to 4 years	5 to 9+ years
Prob. of Employment	-0.138	0.209	-1.009**	-0.302***
	(0.195)	(0.226)	(0.315)	(0.086)
N	2,099,147	344,156	2,075,913	1,083,435
Wages	-0.479	-0.664*	-0.556	-0.194
	(0.344)	(0.279)	(0.354)	(0.182)
N	1,094,691	240,058	1,170,322	785,839
Clusters	109	109	109	109

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

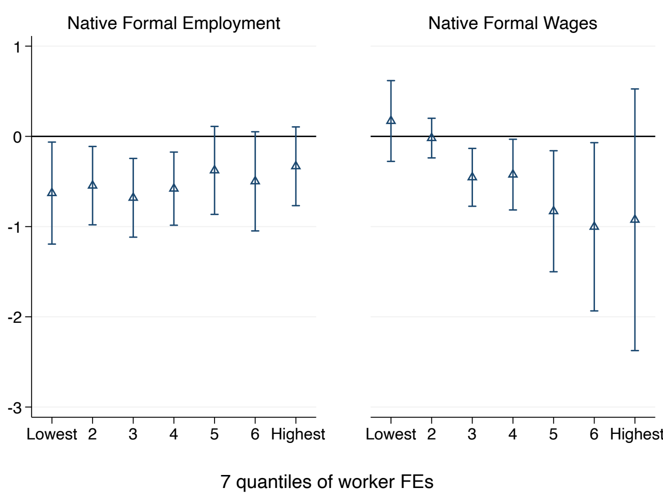
Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to native employees between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2013–2018.

Figure B.2: **Decomposition of formal employment, 2015–2018**



Note: Regressions are estimated at the regional level for 109 FUAs weighted by their formal employment in 2015. The sample is restricted to natives that are self-employed and employees of all ages. Overall regional formal employment is decomposed into outflows from formal employment in that region, inflows from non-employment or the informal sector, and employed people in other regions, and relocation of formal workers to other regions. Source: PILA, 2015–2018.

Figure B.3: **Estimates by quantiles of worker FEs, 2015–2018**



Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to native employees between 25 and 55 years old that appear more than once in PILA. Worker FEs are computed in the first step using the standard AKM framework, with age squared and its cubic as time-varying controls, for the period 2010-2015. Controls used in the second step are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). The coefficients for employment (in percentage points) and for wages (in percent) are already multiplied by 100. Workers are observed in August of each year. 95% confidence interval. Source: PILA, 2013–2019.

Table B.2: **Employment and wage estimates by firm size and age of firm, 2015–2018**

Firm's size	1 to 19 workers		Above 19 workers	
Age of firm	0 to 4 years	5 to 9+ years	0 to 4 years	5 to 9+ years
Prob. of Employment	-0.762**	-0.757***	-1.015**	-0.176
	(0.279)	(0.156)	(0.347)	(0.170)
N	479,715	498,842	923,272	3,700,822
Wages	-1.021*	-0.554	-0.603*	-0.395
	(0.432)	(0.305)	(0.304)	(0.286)
N	274,728	352,015	444,586	2,219,581
Clusters	109	109	109	109

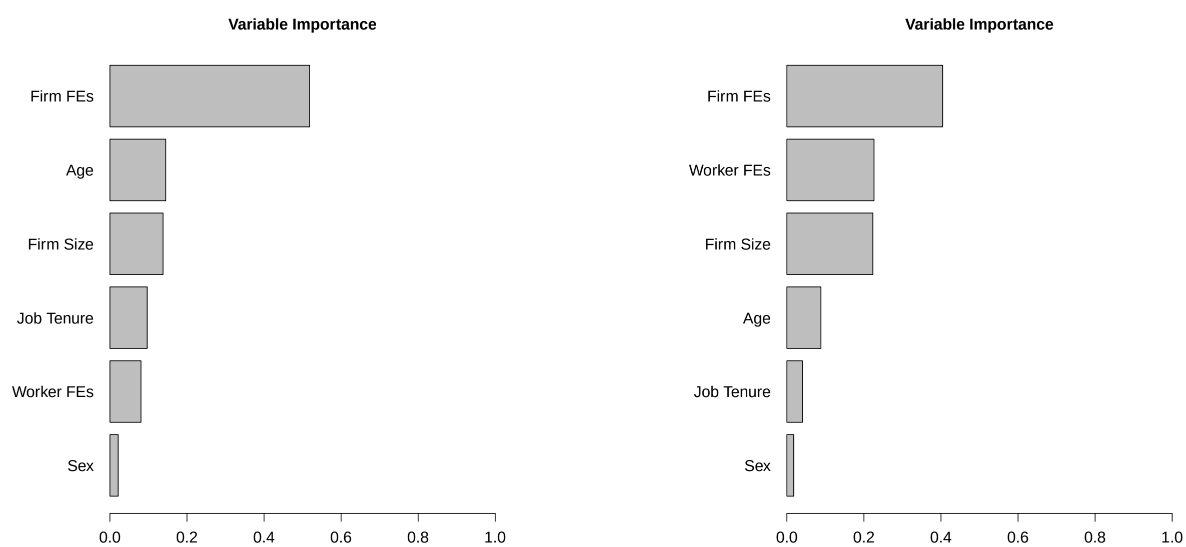
Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Dependent variables are employment relative to the pre-shock period and wages relative to the base period. The sample is restricted to native employees between 25 and 55 years old. The age of the firm is the number of years the firm appears discontinuously in PILA Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2013–2018.

Figure B.4: **Variable importance for formal employment and formal wages in causal forest with worker and firm FEs, 2015–2018**

(a) Employment

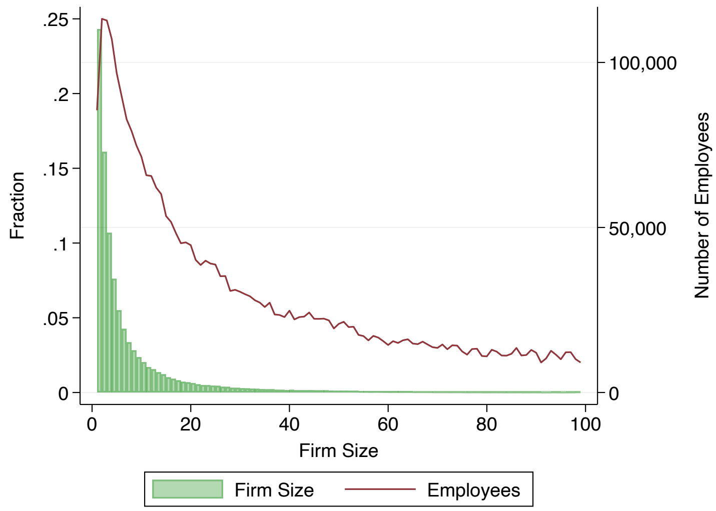
(b) Wages



Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. I use clusters for the causal forest estimation. The minimum node size is 300.

C Robustness Checks

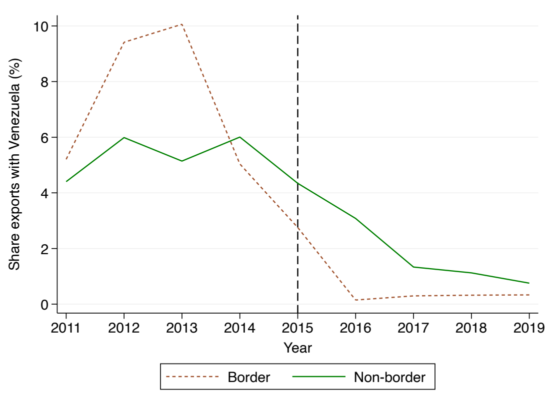
Figure C.1: **Firm size distribution and total employees**



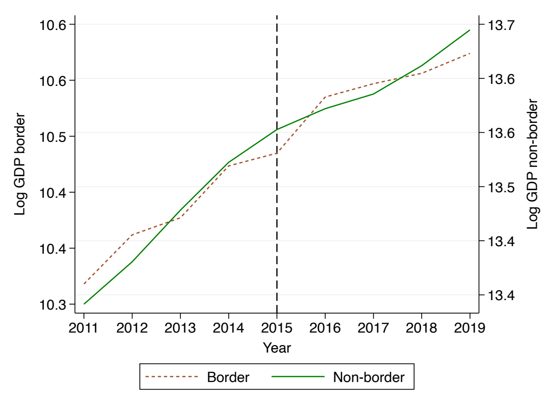
Note: The upper bound of firm size is restricted to 100 workers for the figure. Chosen bin width is 1. Only workers who contribute as employees are taken into account. Source: PILA, August 2015.

Figure C.2: **Evolution of trade and GDP for border and non-border departments**

(a) Share Exports with Venezuela



(b) Log GDP



Note: Border departments are Norte de Santander, La Guajira, and César. Non-border departments are the rest. Source: Panel (a) Exportaciones-DANE, 2013–2019. Panel (b) DANE-Cuentas Nacionales, 2011–2019.

Table C.1: **Robustness checks for formal wages and formal employment, 2015–2018**

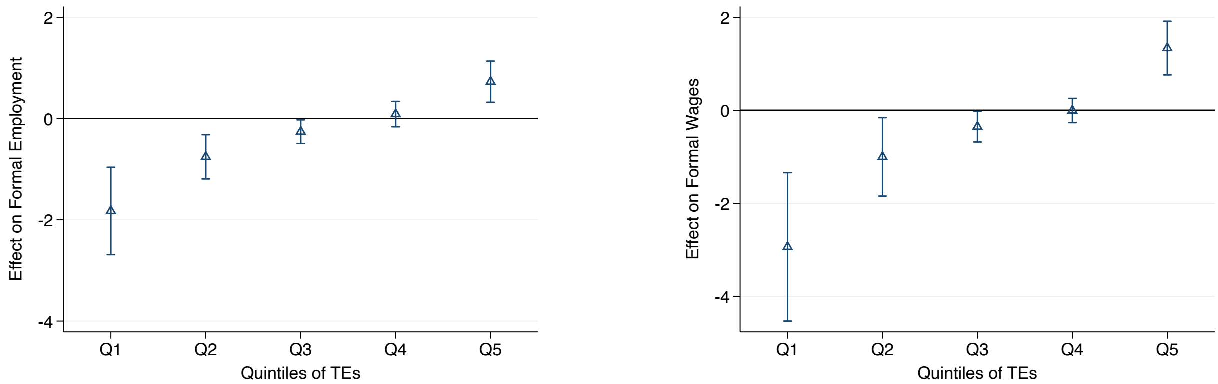
	Employment	Wages
Baseline	-0.841***	-0.600*
	(0.192)	(0.239)
N	6,706,035	4,090,973
Removing border areas*	-1.019*	-0.768
	(0.414)	(0.559)
N	6,577,923	4,015,648
Removing Bogotá	-0.777***	-0.470**
	(0.180)	(0.173)
N	4,338,192	2,619,237
Further controls⋆	-0.484*	-0.556
	(0.191)	(0.286)
N	4,884,993	3,217,398
Real wages		-0.520*
		(0.207)
N		4,090,973
Top code local wages above 99%		-0.605*
		(0.241)
N		4,090,973
Clusters	109	109

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table reports the coefficients of the second-stage regression of the instruments with the immigration rate $\Delta M_{l,2018}$. The outcome is the difference with the base period. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. *The border areas are Cucutá, Maicao and Arauca. ⋆Further controls refer to FEs of seven wage quantiles and job tenure, omitting self-employed workers. The sample is restricted to natives between 25 and 55 years old. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2015–2018.

Figure C.3: **Quintiles of treatment effects for formal employment and formal wages in the causal forest, 2015–2018**

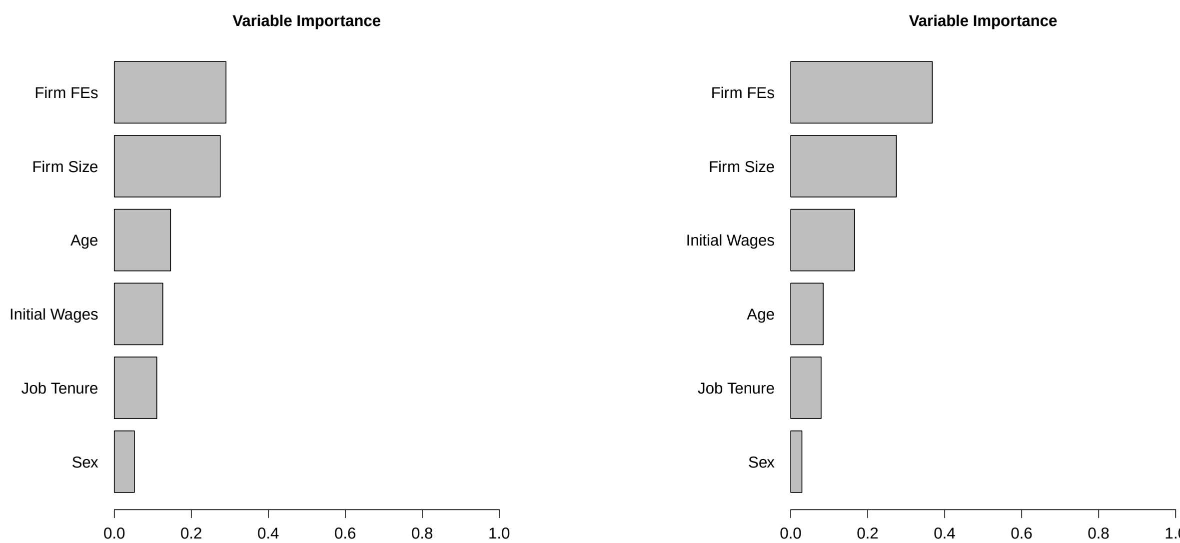
(a) Employment (b) Wages



Note: The quintiles of treatment effects are constructed using the individual treatment effects from the trained causal forest. The coefficients come from separate regressions of the main empirical specification. The sample is restricted to natives between 25 and 55 years old. I use clusters at the FUA level for the causal forest. The causal forest uses 50% of the main sample due to computational burden. The coefficients for employment (in percentage points) and for wages (in percent) are already multiplied by 100. The confidence interval is at the 95% level.

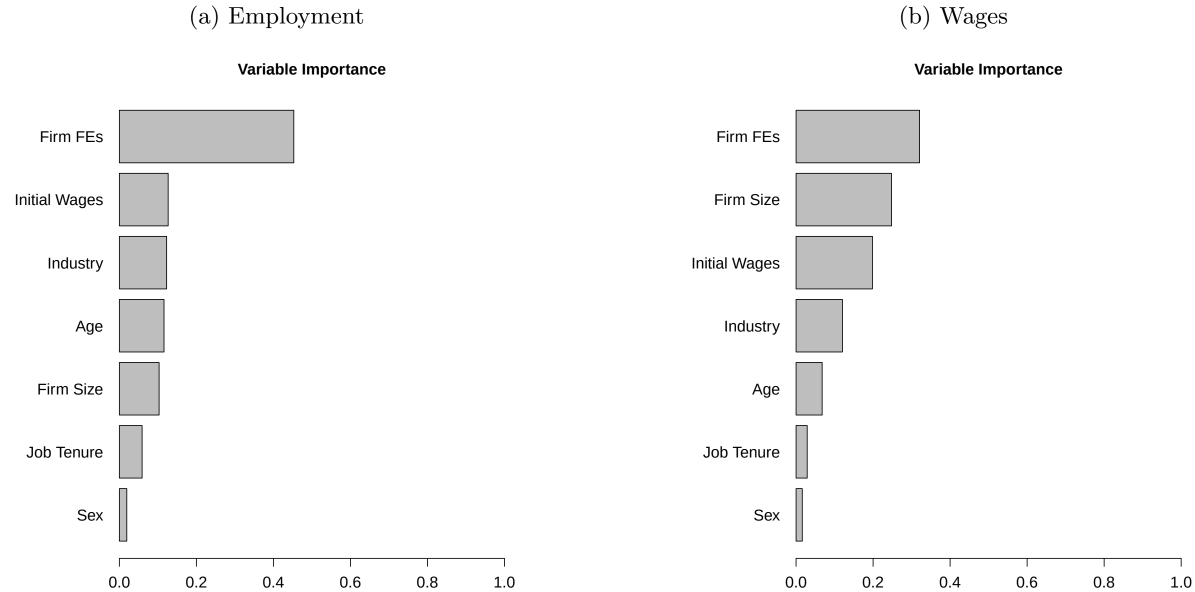
Figure C.4: **Variable importance for formal employment and formal wages in the causal forest with decay exponent, 2015–2018**

(a) Employment (b) Wages



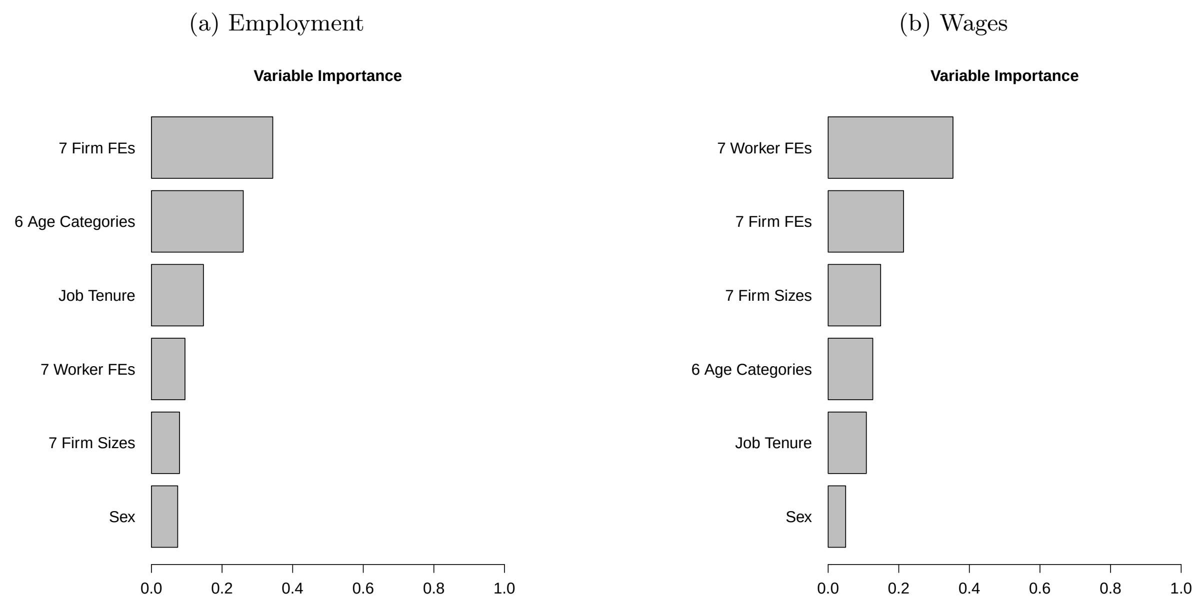
Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. The decay exponent is -2. I use clusters for the causal forest estimation. The minimum node size is 300.

Figure C.5: **Variable importance for formal employment and formal wages in causal forest with industry, 2015–2018**



Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. Industry information is aggregated in 16 industries. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. The decay exponent is -2. I use clusters for the causal forest estimation. The minimum node size is 300.

Figure C.6: **Variable importance for formal employment and formal wages in causal forest for categories, 2015–2018**



Note: Variable importance is a weighted sum of how many times the feature f appears in the split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 25 and 55 years old. The importance measure sums up to 1. I use clusters for the causal forest estimation. The minimum node size is 300.

D Construction of AKM sample

To construct the sample for the AKM estimation, I restrict it to six years before the immigration shock to capture more movements of workers between firms. This sample uses the years 2010 to 2015 for August. The total sample consists of 32,195,048 worker-year observations after eliminating workers with non-positive wages, with less than 30 employment days per month, restricting to employees between 20 and 60 years, and leaving the highest wage job for workers with more than one contribution to the social security system.[D.1] Also, I eliminate 3,931,843 additional workers because they do not belong to the largest connected set of firms and workers or appear only once in the estimation sample. Then, the nominal wages are transformed to real terms using the monthly CPI from DANE (with the base year 2018) and use logarithms of the final expression (lnw_{it}). Table D.1 shows descriptive statistics by the seven quantiles of firm FEs and Table D.2 shows the decomposition of the variance of wages $Var(lnw_{it})$.

Table D.1: **Descriptive statistics by firm FEs**

	Average				
7 quantiles of $\hat{\psi}_j$	Employment	Male (%)	Age	Real wages (USD)	N
1	8	0.6	37.7	239.2	40,201
2	18	0.7	37.1	224.0	41,628
3	14	0.6	37.2	232.5	37,703
4	13	0.6	37.5	248.3	36,223
5	18	0.5	38.0	276.4	36,599
6	40	0.5	38.3	342.0	38,524
7	81	0.5	38.4	616.1	42,455

Note: This table reports the descriptive statistics for different firm sizes recorded in PILA. Real wages are deflated using the CPI from DANE for prices in 2018. Colombian pesos to USD using 2020 exchange rates from the World Bank. Only workers who contribute as employees are taken into account. Source: PILA, August 2015.

Table D.2: Variance decomposition of lnw_{it}

Share of variance explained by:	
$Var(\alpha_i)$	50.2%
$Var(\psi_{j(i)})$	15.7%
$2Cov(\alpha_i, \psi_{j(i)})$	21.6%
$Corr(\alpha_i, \psi_{j(i)})$	.38

Note: This Table reports the variance decomposition of wages in the formal sector in Colombia using the largest connected set of workers and firms with the leave-out method proposed in Kline et al. (2020) with year FEs as the control variable. Source: PILA, August 2010–August 2015.

[D.1] Around 5% of workers in PILA have more than one contribution.

E Derivations of Model in subsection 5.1

In this Appendix section, I explain the derivations of the equations in subsection 5.1. First, to derive the firm-specific optimal wages, I maximize the profit equation (9) for each type of worker[E.1]:

$$\frac{d\pi_j}{dI_j} = 0 \Leftrightarrow w_{I_j} = \left(\frac{\beta_I(1+\eta)}{1+\beta_I(1+\eta)}\right) D_j T_j^{\epsilon} \epsilon \alpha_I I_j^{\rho-1-\eta}(1+\eta)^{-1}(\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho})^{\frac{\epsilon-\rho}{\rho}}, \tag{E.1}$$

$$\frac{d\pi_j}{dF_j} = 0 \Leftrightarrow w_{F_j} = \left(\frac{\beta_F}{1+\beta_F}\right) D_j T_j^{\epsilon} \epsilon \alpha_F F_j^{\rho-1}(1+\tau_F)^{-1}(\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho})^{\frac{\epsilon-\rho}{\rho}}. \tag{E.2}$$

Here, workers' wages not only depend on their marginal productivity but also on the size of the labor supply elasticities to the firm.[E.2] Importantly, if there exists a minimum wage for formal workers ($w_{F_{Min}}$) such that $w_{F_{Min}} \leq w_{F_j}$, then formal workers must be paid the minimum wage and firms' optimal labor choices would be distorted. In general, this is more likely to happen in low-productivity firms. Moreover, firms with higher productivity or demand will pay higher wages, holding constant amenities. For clarity, l take logarithms of the wages in equation (E.1) and (E.2):

$$lnw_{I_j} = ln\left(\frac{\beta_I(1+\eta)}{1+\beta_I(1+\eta)}\right) + ln(D_j T_j^{\epsilon} \epsilon \alpha_I) + (\rho-1-\eta) lnI_j - ln(1+\eta) + \left(\frac{\epsilon-\rho}{\rho}\right) ln(\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho}), \tag{E.3}$$

$$lnw_{F_j} = ln\left(\frac{\beta_F}{1+\beta_F}\right) + ln(D_j T_j^{\epsilon} \epsilon \alpha_F) + (\rho-1) lnF_j - ln(1+\tau_F) + \left(\frac{\epsilon-\rho}{\rho}\right) ln(\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho}). \tag{E.4}$$

Next, I study how firm-specific wages respond to an immigration shock that shifts the aggregate informal labor supply outwards ($d\mathcal{I}$)[E.3]:

$$\frac{dlnw_{I_j}}{d\mathcal{I}} \cdot \mathcal{I} = (\rho-1-\eta)\frac{dlnI_j}{dln\mathcal{I}} + \left(\frac{\epsilon-\rho}{\rho}\right)\frac{(\alpha_I \rho I_j^{\rho-1}\frac{dI_j}{d\mathcal{I}} + \alpha_F \rho F_j^{\rho-1}\frac{dF_j}{d\mathcal{I}})}{\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho}} * \mathcal{I}, \tag{E.5}$$

$$\frac{dlnw_{F_j}}{d\mathcal{I}} \cdot \mathcal{I} = (\rho-1)\frac{dlnF_j}{dln\mathcal{I}} + \left(\frac{\epsilon-\rho}{\rho}\right)\frac{(\alpha_I \rho I_j^{\rho-1}\frac{dI_j}{d\mathcal{I}} + \alpha_F \rho F_j^{\rho-1}\frac{dF_j}{d\mathcal{I}})}{\alpha_I I_j^{\rho} + \alpha_F F_j^{\rho}} * \mathcal{I}. \tag{E.6}$$

Simplifying the last expressions and defining the derivates as the elasticities, I find that:

$$\varepsilon_{w_{I_j},\mathcal{I}} = -(1+\eta-\rho)\varepsilon_{I_j,\mathcal{I}} + (\epsilon-\rho)(s_{I_j}\varepsilon_{I_j,\mathcal{I}} + s_{F_j}\varepsilon_{F_j,\mathcal{I}}), \tag{E.7}$$

[E.1] In the derivations, I multiply by $\frac{w(L_j)}{w(L_j)}$ in the last term of FOCs to find the equations on the text.

[E.2] If $\beta_L = 9$ then workers are paid 90% of their marginal productivity to the firm.

[E.3] Assuming that the supply shock does not affect the firm-specific demand and the firm-specific amenities for each group of workers. Besides, the number of firms is sufficiently large such that there are no strategic interactions between firms.

$$\varepsilon_{w_{F_j},\mathcal{I}} = -(1-\rho)\varepsilon_{F_j,\mathcal{I}} + (\epsilon-\rho)(s_{I_j}\varepsilon_{I_j,\mathcal{I}} + s_{F_j}\varepsilon_{F_j,\mathcal{I}}). \tag{E.8}$$

In these expressions, $s_{L_j} = \frac{\alpha_L L_j^\rho}{\alpha_I I_j^\rho + \alpha_F F_j^\rho}$ is the relative contribution of type of worker $L \in \{I, F\}$ to production. Next, I derive the changes from the immigration shock using the firm-specific supply functions (7) and (8):

$$\varepsilon_{I_j,\mathcal{I}} = 1 + \beta_I \varepsilon_{w_{I_j},\mathcal{I}}, \tag{E.9}$$

$$\varepsilon_{F_j,\mathcal{I}} = \beta_F \varepsilon_{w_{F_j},\mathcal{I}}. \tag{E.10}$$

This yields a direct relationship between wages and employment as a function of the elasticities of supply to the firm.[E.4] Then, l replace equations (E.9) and (E.10) into (E.7) and into (E.8):

$$\varepsilon_{w_{I_j},\mathcal{I}} = -(1+\eta-\rho)(1+\beta_I\varepsilon_{w_{I_j},\mathcal{I}}) + (\epsilon-\rho)(s_{I_j}(1+\beta_I\varepsilon_{w_{I_j},\mathcal{I}}) + s_{F_j}\beta_F\varepsilon_{w_{F_j},\mathcal{I}}), \tag{E.11}$$

$$\varepsilon_{w_{F_j},\mathcal{I}} = -(1-\rho)\beta_F\varepsilon_{w_{F_j},\mathcal{I}} + (\epsilon-\rho)(s_{I_j}(1+\beta_I\varepsilon_{w_{I_j},\mathcal{I}}) + s_{F_j}\beta_F\varepsilon_{w_{F_j},\mathcal{I}}). \tag{E.12}$$

Rearranging these expressions I find that:

$$\varepsilon_{w_{I_j},\mathcal{I}} = \left(\frac{1}{\xi_{I_j}}\right)(-(1+\eta-\rho) + (\epsilon-\rho)(s_{I_j} + s_{F_j}\beta_F\varepsilon_{w_{F_j},\mathcal{I}})), \tag{E.13}$$

$$\varepsilon_{w_{F_j},\mathcal{I}} = \left(\frac{1}{\xi_{F_j}}\right)(\epsilon-\rho)s_{I_j}(1+\beta_I\varepsilon_{w_{I_j},\mathcal{I}}). \tag{E.14}$$

Here, I define $\xi_{I_j} = 1 + (1+\eta-\rho)\beta_I - (\epsilon-\rho)s_{I_j}\beta_I$ and $\xi_{F_j} = 1 + (1-\rho)\beta_F - (\epsilon-\rho)s_{F_j}\beta_F$. Then, replacing equation (E.13) into (E.14) yields:

$$\varepsilon_{w_{F_j},\mathcal{I}} = \Omega_j s_{I_j}\beta_I(\epsilon-\rho)(\frac{\xi_{I_j}}{\beta_I} - (1+\eta-\rho) + (\epsilon-\rho)s_{I_j}). \tag{E.15}$$

Here, I define $\Omega_j = \frac{1}{\xi_{I_j}\xi_{F_j} - (\epsilon-\rho)^2 s_{I_j}\beta_I s_{F_j}\beta_F}$. Last, I replace ξ_{I_j} inside of (E.15) to find the equation (10) in the main text. Next, I plug equation (10) inside equation (E.13) to find that:

$$\varepsilon_{w_{I_j},\mathcal{I}} = \left(\frac{1}{\xi_{I_j}}\right)(-(1+\eta-\rho) + (\epsilon-\rho)s_{I_j}(1 + s_{F_j}\Omega_j(\epsilon-\rho)\beta_F)). \tag{E.16}$$

In this case, if the firm does not hire any formal workers (i.e., $s_{F_j} = 0$) then $\varepsilon_{w_{I_j},\mathcal{I}} < 0$, and for the rest of values of s_{F_j} this elasticity is also going to be negative.[E.5] Finally, after finding that

[E.4] Here, the total number of formal workers $\mathcal{F}$ in the market is held constant. Besides, in this partial equilibrium framework, the response of one firm does not have spillover effects on other firms.

[E.5] To find that $\varepsilon_{w_{I_j},\mathcal{I}} < 0$ it is sufficient that $1 \geq s_{I_j}(1 + s_{F_j}\Omega_j(\epsilon-\rho)\beta_F)$, which always happens when $\rho > \epsilon$. On informal wages always decrease with the informal supply shock, the last adjustment to analyze is what happens to informal employment within the firm. For that, I plug equation (E.16) into equation (E.9):

$$\varepsilon_{I_j,\mathcal{I}} = 1 + \left(\frac{\beta_I}{\xi_{I_j}}\right)(-(1+\eta-\rho) + (\epsilon-\rho)s_{I_j}(1 + s_{F_j}\Omega_j(\epsilon-\rho)\beta_F)). \tag{E.17}$$

After simplifying the previous expression, I find that:

$$\varepsilon_{I_j,\mathcal{I}} = \frac{1}{\xi_{I_j}}(1 + (\epsilon-\rho)^2 s_{I_j}\beta_I s_{F_j}\beta_F\Omega_j). \tag{E.18}$$

Thus, in this case, a positive aggregate informal supply shock always increases informal labor within the firm ($\varepsilon_{I_j,\mathcal{I}} > 0$), independent of formal and informal workers' being close substitutes or not.

F Additional tests for Pre-Trends

This subsection of the Appendix tests for differential trends in the outcomes according to different workers' characteristics.

the other hand, if $\rho < \epsilon$ then $\varepsilon_{w_{I_j},\mathcal{I}} < 0$ is also negative as $1+\eta-\rho > \epsilon-\rho$.

Table F.1: Event study estimates on pre-treatment periods of Figure 5a

	Employment			Wages		
	2012	2013	2014	2012	2013	2014
25 to 30 years	-0.312	-0.458	-0.412	0.284	0.370	-0.019
	(0.702)	(0.655)	(0.428)	(0.461)	(0.234)	(0.124)
30 to 35 years	-0.249	-0.512	-0.305	0.104	0.226	-0.011
	(0.499)	(0.415)	(0.322)	(0.166)	(0.235)	(0.272)
35 to 40 years	-0.212	-0.381	-0.060	0.032	0.166	-0.011
	(0.364)	(0.327)	(0.196)	(0.296)	(0.253)	(0.273)
40 to 45 years	0.043	-0.155	-0.328	-0.012	-0.068	-0.501**
	(0.367)	(0.318)	(0.240)	(0.297)	(0.275)	(0.160)
45 to 50 years	0.191	-0.092	-0.101	0.110	0.566	-0.032
	(0.303)	(0.266)	(0.215)	(0.290)	(0.356)	(0.187)
50 to 55 years	0.121	0.005	0.103	0.413	0.369	-0.209
	(0.335)	(0.262)	(0.191)	(0.293)	(0.313)	(0.201)
Males	-0.449	-0.715	-0.473	0.272	0.317	-0.018
	(0.512)	(0.481)	(0.322)	(0.235)	(0.285)	(0.170)
Females	0.297	0.193	0.094	-0.011	0.199	-0.225*
	(0.376)	(0.310)	(0.223)	(0.209)	(0.178)	(0.095)

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The sample is reduced to a 10% random subsample of the entire dataset due to computational burden. The sample is restricted to natives between 25 and 55 years old. Controls used are sex and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2012–2015.

Table F.2: Event study estimates on pre-treatment periods of Figure 7a

	Employment			Wages		
	2012	2013	2014	2012	2013	2014
Minimum wage	0.313	0.296	0.217	-0.538	-0.409	-0.526
	(0.269)	(0.213)	(0.181)	(0.825)	(0.700)	(0.435)
40th–50th	-0.295	-0.886	-0.190	0.262	0.190	-0.107
	(0.502)	(0.471)	(0.403)	(0.388)	(0.449)	(0.235)
50th–60th	-0.235	-0.766*	-0.284	0.058	0.234	-0.141
	(0.469)	(0.382)	(0.240)	(0.335)	(0.298)	(0.170)
60th–70th	-0.244	-0.100	-0.136	-0.360	0.401*	0.150
	(0.319)	(0.321)	(0.226)	(0.264)	(0.186)	(0.120)
70th–80th	-0.243	-0.475	-0.553**	0.918	0.730*	0.033
	(0.301)	(0.281)	(0.211)	(0.481)	(0.341)	(0.259)
80th–90th	-0.130	-0.432	-0.385*	0.435	0.367	-0.026
	(0.330)	(0.241)	(0.167)	(0.596)	(0.485)	(0.268)
90th–100th	0.330	-0.220	-0.146	-0.039	0.132	-0.477
	(0.483)	(0.173)	(0.136)	(0.288)	(0.269)	(0.297)

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The sample is reduced to a 10% random subsample of the entire dataset due to computational burden. The sample is restricted to natives between 25 and 55 years old. Controls used are interactions of sex with six age categories and a dummy for self-employed in the base period. Cluster standard errors (G=109). Workers are observed in August of each year. Source: PILA, 2012–2015.

G Information of FUAs

Table G.1: Number of observations by FUA I

	Observations	Percent
1. Bogotá	2,327,306	(32.7)
2. Medellín	983,096	(13.8)
3. Cali	593,447	(8.3)
4. Barranquilla	341,211	(4.8)
5. Cartagena	205,150	(2.9)
6. Bucaramanga	273,090	(3.8)
7. Cúcuta	110,123	(1.5)
8. Pereira	140,791	(2.0)
9. Ibagué	100,823	(1.4)
10. Manizales	103,401	(1.5)
11. Santa Marta	84,705	(1.2)
12. Pasto	70,170	(1.0)
13. Armenia	71,314	(1.0)
14. Villavicencio	106,493	(1.5)
15. Montería	71,007	(1.0)
16. Valledupar	76,072	(1.0)
17. Buenaventura	24,514	(0.3)
18. Neiva	71,376	(1.0)
19. Palmira	41,687	(0.6)
20. Popayán	62,422	(0.9)
21. Sincelejo	39,859	(0.6)
22. Barrancabermeja	35,095	(0.5)
23. Tuluá	25,123	(0.3)
24. Tunja	52,987	(0.7)
25. Riohacha	31,134	(0.4)
26. San Andres de Tumaco	7,960	(0.1)
27. Florencia	19,704	(0.3)
28. Apartadó	26,268	(0.4)
29. Giradot	14,920	(0.2)
30. Cartago	17,006	(0.2)
31. Maicao	6,263	(0.1)
32. Magangué	5,327	(0.1)
33. Sogamoso	18,220	(0.3)
34. Buga	21,072	(0.3)
35. Ipiales	8,754	(0.1)
36. Quibdó	15,687	(0.2)
37. Fusagasugá	12,899	(0.2)
38. Facatativá	18,796	(0.3)
39. Duitama	18,427	(0.3)
40. Yopal	43,279	(0.6)
41. Ciénaga	4,701	(0.1)
42. Zipaquirá	12,908	(0.2)
43. Rionegro	29,601	(0.4)
44. Ocaña	8,966	(0.1)
45. La Dorada	8,563	(0.1)
46. Caucasia	7,372	(0.1)
47. Sabanalarga	2,434	(0.03)
48. Aguachica	9,748	(0.1)
49. Espinal	6,439	(0.1)
50. Arauca	11,726	(0.2)
51. Santa Rosa de Cabal	4,887	(0.1)
52. El Carmen de Bolívar	1,411	(0.02)
53. Fundación	3,881	(0.1)
Continues in Table G.2		
No FUA assigned	417,188	(5.9)
Total	7,123,223	(100)

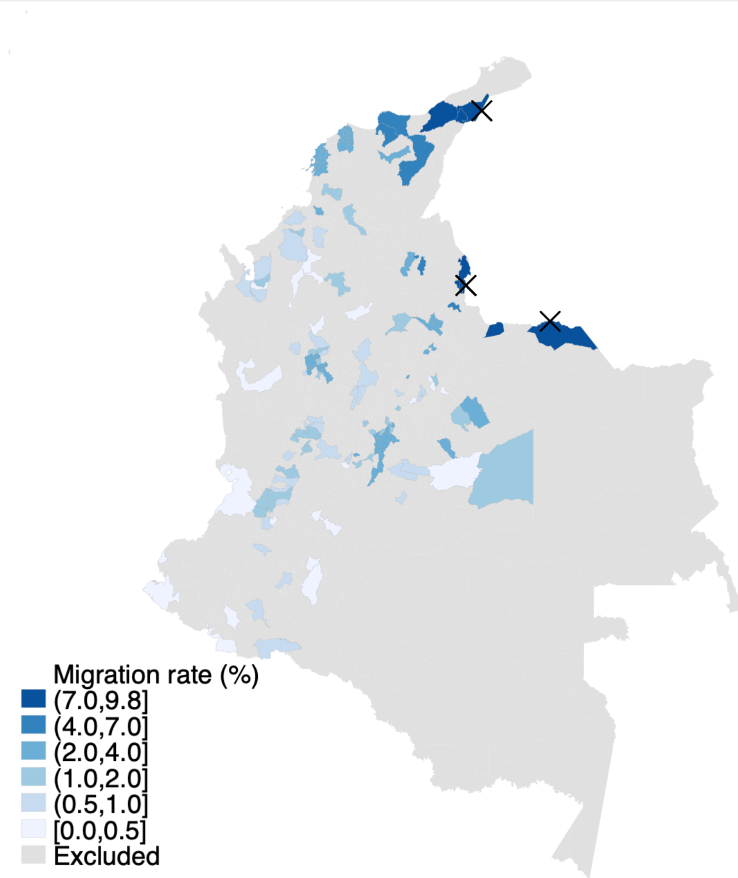
Note: This Table reports the number of workers from PILA by FUAs 1 to 53. The name represents the main city of FUA but often they aggregate multiple municipalities according to Sanchez-Serra (2016). The sample is restricted to natives between 25 and 55 years old. Workers are observed in August of each year. Source: PILA, 2015.

Table G.2: Number of observations by FUA II

	Observations	Percent
54. Acacías	12,472	(0.2)
55. Madrid	8,922	(0.1)
56. La Ceja	8,662	(0.1)
57. Santander de Quilichao	8,505	(0.1)
58. San Gil	8,268	(0.1)
59. Mocoa	7,974	(0.1)
60. Pitalito	7,852	(0.1)
61. Albania	7,020	(0.1)
62. Tocancipá	7,007	(0.1)
63. Los Patios	6,137	(0.1)
64. Montelíbano	6,083	(0.1)
65. Turbo	5,830	(0.1)
66. Granada	5,298	(0.1)
67. El Carmen de Viboral	5,047	(0.1)
68. Chinchiná	4,903	(0.1)
69. Puerto Boyacá	4,761	(0.1)
70. Guarne	4,697	(0.1)
71. Zarzal	4,584	(0.1)
72. Puerto Asís	4,568	(0.1)
73. Chiquinquirá	4,526	(0.1)
74. Villa de San Diego de Ubaté	4,522	(0.1)
75. Garzón	4,454	(0.1)
76. Santa Rosa de Osos	4,406	(0.1)
77. Puerto Gaitán	4,380	(0.1)
78. Pamplona	4,348	(0.1)
79. Puerto Tejada	4,279	(0.1)
80. Caloto	4,136	(0.1)
81. Segovia	4,016	(0.1)
82. Puerto Berrío	3,989	(0.1)
83. Lorica	3,875	(0.1)
84. Sopó	3,832	(0.1)
85. Aguazul	3,627	(0.1)
86. Santa Fé de Antioquia	3,589	(0.1)
87. Cereté	3,526	(0.0)
88. Puerto López	3,412	(0.0)
89. Pradera	3,388	(0.0)
90. La Cruz	3,387	(0.0)
91. La Virginia	3,375	(0.0)
92. San Pedro de los Milagros	3,170	(0.0)
93. Tenjo	3,166	(0.0)
94. Villanueva	3,136	(0.0)
95. Sahagún	3,126	(0.0)
96. Melgar	3,099	(0.0)
97. Barbosa, Santander	3,042	(0.0)
98. Socorro	3,026	(0.0)
99. Carepa	2,999	(0.0)
100. Planeta Rica	2,893	(0.0)
101. Chigorodó	2,880	(0.0)
102. Yarumal	2,874	(0.0)
103. Paipa	2,873	(0.0)
104. Samacá	2,782	(0.0)
105. Barbosa, Antioquia	2,781	(0.0)
106. Saravena	2,730	(0.0)
107. El Cerrito	2,597	(0.0)
108. Amagá	2,534	(0.0)
109. Villeta	2,518	(0.0)

Note: This Table reports the number of workers from PILA by FUAs 54 to 109. The name represents the main municipality. The sample is restricted to natives between 25 and 55 years old. Workers are observed in August of each year. Source: PILA, 2015.

Figure G.1: **Map of FUAs with the migration shock** $\Delta M_{l,2018}$



Note: The X represents the main three crossing bridges with Venezuela. The distance instrument is according to the nearest crossing bridge. Source: CNPV, 2018.

G.1 Definition of Variables

Formal wages. The nominal contribution to the health system of each worker for August is used. Only positive contributions are considered, as zero indicates workers on leave for several reasons unrelated to wages or jobs. The focus is on workers who reported 30 days of employment.
Natives with formal employment. All individuals that appear in PILA with a national identity card are counted as natives. All the natives in the sample with a non-negative wage are taken as employed.
Firms. I only leave workers classified as employees for the firm-level data and then aggregate by the firm identifier.