The Role of Workers and Firms in the Impact of Immigration*

Lukas Delgado-Prieto†

May , 2022

Abstract

This paper studies the worker-level effects of a labor supply shock and determines the differential role of workers and firms in these effects. To do so, I exploit Venezuelans' uneven and massive arrival within Colombia (as of 2019, nearly 2 million Venezuelans lived in Colombia) and use administrative employer-employee data covering the universe of formal workers to follow natives' labor market outcomes over time. Overall, I find a reduction in worker-level employment that is concentrated at the bottom of the wage distribution (among self-employed and minimum wage earners). In contrast, I find a negative wage effect that is driven by workers from the upper part of the wage distribution who work in relatively small firms. To uncover the mechanisms behind these findings, I construct a vector of observed and unobserved worker and firm characteristics and implement a machine learning method. This method shows that firm-specific pay premiums are more important in explaining the observed drop in employment and wages than other worker characteristics. These results support the influential role that firms play in determining the impact of immigration on workers' outcomes.

Keywords: Immigration, Labor market, Causal forest.

JEL Codes: F22, O15, O17, R23.

*I am grateful to Jan Stuhler, Luigi Minale, Juanjo Dolado, Jaime Arellano-Bover and Jesús Fernández-Huertas for their comments and suggestions. I am also thankful for feedback at the UC3M Applied Reading Group.

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

1 Introduction

In this paper, I use information from the universe of formal workers in Colombia to study the labor market impacts of one of the most significant episodes of immigration in recent history, the Venezuelan mass migration to Colombia.[1] Exploiting the unequal arrival of Venezuelan immigrants across local labor markets, combined with an instrumental variable (IV) approach, I quantify the worker-level effects of immigration and the role that workers and firms have in these effects. Because I can follow workers over time, is possible to account for compositional changes that are not identified when using cross-sectional data for the analysis of immigration effects.

To my knowledge, this is one of the first papers to study labor supply shocks in developing countries equipped with panel administrative data.[2] This allows to estimate a rich set of heterogeneous effects by the worker and firm characteristics and to implement a machine learning method that help to explain the underlying mechanisms of immigration adjustments.

A key characteristic of Latin American labor markets is the interdependence between the informal and formal sectors, which translates into asymmetric employment and wage responses across these sectors in the face of immigration shocks (as discussed in Kleemans and Magruder (2018) for Indonesia and Delgado-Prieto (2022) for Colombia). Specifically, Delgado-Prieto (2022) find a large negative effect on formal regional employment from the Venezuelan immigration that motivates the individual analysis undertaken here.

My empirical strategy compares similar workers in areas with different exposure to immigration over time. Because migrants endogenously sort themselves into areas that offer the best economic opportunities, I use past settlements of Venezuelans and distance to the nearest crossing bridge with Venezuela as instruments. In terms of results, I find a persistent negative impact on individual employment and wages for natives. The negative employment effect is driven by workers earning the minimum wage, above 35 years old, and that were self-employed before the immigration shock. In contrast, the negative wage effect mainly affects native workers earning above the minimum wage and working in the smallest firms.

Next, I exploit the canonical Abowd et al. (1999) (AKM hereafter) framework to recover firm

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

[2]In developed countries, Dustmann et al. (2017) for Germany, Arellano-Bover and San (2020) for Israel, Foged and Peri (2016) for Denmark and Orefice and Peri (2020) for France have recently studied immigration shocks exploiting administrative data.

effects, or firm-specific pay premiums, and worker effects, or worker-specific pay premiums. A significant contribution of this study is to understand the sources of wage and employment losses stemming from immigration using these constructed variables. I find that workers in high-paying firms at the base period are suffering the largest wage losses compared to workers in low-paying firms. A possible explanation is the reallocation effects of immigration, that is, workers might be moving from high- to low-paying firms. However, the results suggest no differential sorting of native workers after migrants arrive. Hence, there seems to be lower wage growth within high-paying firms.[3] Regarding employment, the finding is the opposite: workers in high-paying firms experience more minor impacts than workers in low-paying firms. A similar picture is found when dividing workers by their worker-specific pay premiums.

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 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. The selection of these variables relies on the frequency that they appear in the splits of all the decision trees. In this exercise, I consistently find that firm-specific pay premiums are more important in determining the employment and wage losses than other worker characteristics (i.e., job tenure, age, gender, and wages) in the base period. Therefore, firms' role in the impact of immigration appears to be more important than workers' role, which is one of the main findings of this paper. These findings suggest direct policy implications. For instance, to alleviate the largest losses of immigration, specific firm-level policies might be determined depending on the pay premiums of the firm.

Concerning the related literature, few papers exploit longitudinal administrative data to study migration shocks. The first paper by Foged and Peri (2016) exploits a refugee dispersal policy in Denmark to find that low-skilled natives pursue less manual-intensive occupations, upgrading their wages. The second paper by Dustmann et al. (2022b) discusses the labor market effects of immigration on regions and workers exploiting a commuting policy in Germany. In a different setup, Orefice and Peri (2020) uses a matched employee-employer dataset for France to study the impact of

[3] If there are lower outside options for natives in high-paying firms after migrants arrive. This might increase the bargaining power of these firms and incentivize lower wage growth for its workers.

immigration on worker-firm sorting. This paper finds a causal increase in the assortative matching, with high-quality firms attracting high-quality workers. Another related paper by Arellano-Bover and San (2020) studies the role of firms on the assimilation of immigrants in the labor market. Exploiting the arrivals of former Soviet Union Jews to Israel in the 1990s, these authors find that immigrants start their careers in small and low-paying firms, but over time, they move to larger and higher-paying firms. Concerning the trade literature, Autor et al. (2014) study the effect of trade shocks at the worker-level in the US. Using industry shocks to import competition from China, these authors find a negative impact on earnings, especially for low-wage workers with lower job tenure. Using labor demand shocks, Yagan (2019) analyze the impact of local unemployment shocks arising from the great recession on individual wages and employment of workers in the US, finding permanent employment and wage losses. Last, Gulyas et al. (2019) exploits mass layoffs in Austria, combined with three decades of administrative data, to find that cumulative earning losses are higher among workers in high-paying firms. In addition, Gulyas et al. (2019) is one of the first papers to implement the machine learning algorithm in Athey et al. (2019) to understand the sources of earnings losses after job displacement.

This paper contributes to different strands of the literature. The first contribution is to estimate the individual impact of immigration (see Dustmann et al. (2022b) for the main advantages of it). This is conceptually different from the standard analysis at the regional-level (for instance, see Dustmann et al. (2017) for Germany, Monras (2020) for US and Aksu et al. (2018) for Turkey). By following workers over time, I can integrate into the analysis all the movements of natives between areas, reducing the attenuation of the estimates discussed in Borjas (2006). The second one relates to the impact of immigration in developing countries (see related papers Caruso et al. (2021); Delgado-Prieto (2022); Lebow (2021); Morales-Zurita et al. (2020); Santamaria (2020); Peñaloza Pacheco (2019) for Colombia, Kleemans and Magruder (2018) for Indonesia, or Aksu et al. (2018); Ceritoglu et al. (2017); Del Carpio and Wagner (2015) for Turkey). Since it is unusual to have administrative data in developing countries, all the previous studies used surveys to determine the impact of immigration. Therefore, using panel administrative data, I can quantify detailed worker-level effects of the impact of immigration in a developing country. The third one is identifying the main drivers of labor market adjustments to a large immigration shock. Equipped with the universe of formal firms and workers, I can determine the sources of these adjustments using a machine learning algorithm (similar to Gulyas et al. (2019) for mass layoffs). Usually, heterogeneous effects are derived from the interaction of individual characteristics with the treatment or by selecting groups of the population, but with the machine learning method I determine the variables that explains most of the heterogeneity of immigration effects.

The rest of the paper is structured as follows. Section 2 discusses the characteristics of the labor supply shock of Venezuelans. Section 3 describes the data used and descriptive statistics. Section 4 describes the empirical strategy and the required identification assumptions. Sections 5 and 6 report results at the worker-level by different individual and firm characteristics. Section 7 introduces the machine learning approach used in the paper and discusses the main findings with this method. Section 8 provides robustness tests. Finally, Section 9 concludes.

2 The Venezuelan Mass Migration

Venezuela's unprecedented socio-economic and political deterioration has triggered massive outflows of people leaving the country since 2016, both in a voluntary and a forced manner. As a result, several countries in Latin America are receiving vast numbers of Venezuelans, especially Colombia, Perú, and Ecuador UNHCR (2019). By far, Colombia has been the major receiver country with more than 1.2 million working-age Venezuelans (4.1% of the working-age population living in Colombia) as of 2019 (DANE, 2019). These sudden inflows can alter different socio-economic outcomes in the short and long-run in the host countries.

The Venezuelan exodus is unmatched in the recent history of migration in Latin America and, worldwide, there are only two contexts that can be comparable in magnitude, namely, the Syrian exodus 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., Aksu et al. (2018), Ceritoglu et al. (2017), or Del Carpio and Wagner (2015)). 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 can get work permits. Yet, around 90% of Venezuelan immigrants were employed in the informal sector 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 levels of education compared to their Colombian counterparts.

As described above, since the labor supply shock in Colombia occurs in the informal sector, why then focus on the formal sector in this paper? The reason is that the immigration shock first affects wages in the informal sector and then affects employment in the formal sector. In effect, the vast majority of migrants work in the unregulated informal labor market, decreasing the cost of informal labor. However, as some firms combine formal and informal labor in their production function (especially in the smallest firms), employers will substitute informal for formal workers (since they are now cheaper) when both types of labor have a high degree of substitutability, as shown in Delgado-Prieto (2022). Therefore, native formal employment is the most affected, even if migrants primarily work in the informal sector. Focusing on formal workers' adjustments, therefore, is of central interest.

3 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, gender, 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. 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 immigration shock. The census provides the most reliable source of information of Venezuelan immigrants.[4]

For the analysis, a dataset is built with all the individuals that appear in PILA, from 2013 to 2019, in the rows and their yearly variables on the columns.[5] The total number of workers

[4]The labor force survey (GEIH, by its acronym in Spanish) also measures the number of Venezuelan immigrants in Colombia at a higher frequency. Nevertheless, there is a measurement error coming from traditional surveys that might attenuate the estimates (Aydemir and Borjas, 2011).

[5]The administrative records of PILA are constructed at the level of the contribution, as workers with more than

in this dataset, which appear at least in one year, is 17,956,372. Next, I restrict to full-time native workers between 20 and 55 years of age in 2015 and assign the immigration shock to all these workers according to their location, which leaves 7,123,192 workers.[6] Then, I transform the municipality variable into a more standard definition of local labor markets or commuting zones following the methodology of Sanchez-Serra (2016).[7] This definition yields 53 functional urban areas (FUAs), after eliminating small or rural municipalities, with a sample of 6,409,615 workers.[8] 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 full employment days in the month and positive wages; moreover, the worker must be employed in the post-treatment year of comparison. Thus, the sample varies slightly year by year (an unbalanced panel). For the impact of immigration by firm characteristics, I eliminate self-employed workers from the sample as they do not belong to a comparable definition of the firm. Is worth noting that all the restrictions to the dataset are the typical ones implemented in the literature.

3.1 Descriptive Statistics for Formal Workers

Table 1a and 1b shows descriptive statistics for natives and migrants by age, gender and wages across time.[9] In terms of observable characteristics, migrants in the formal sector are younger, predominantly male and earn lower wages compared to natives (see Table 1a and 1b). In addition, there is a small share of Venezuelans with the PEP working in the formal sector. It is not possible to observe informal workers in the administrative data, but they represent around half of all workers employed in Colombia.

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. And second, I aggregate the income for workers with more than one labor contract and leave the characteristics only for the job with the highest wage by the worker.

[6] Constructing the treatment in this way this rules out inflows of workers in the post-treatment period from the analysis.

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

[8] The definition of FUAs consists of the 53 biggest urban areas in the country defined from population grid data, municipal boundaries, and inter-municipal commuting flows. In Appendix Table A.3, I show the distribution of the sample by FUAs; using this definition 10% of workers are excluded.

[9] To identify foreign status in PILA, I exploit the type of document workers have in their health contribution. When workers had a national identity card, 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 migrants. Since the PEP's program started around 2018 to foster the regularization of Venezuelan migrants, it is not possible to identify these migrants before that year.

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

(a) Colombians

	Age		Sex (%)		Real wages (USD)		
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	N
2013	37.0	10.8	0.56	0.496	420.3	503.3	7,364,745
2015	37.2	11.1	0.56	0.497	416.8	484.5	8,391,852
2017	37.8	11.4	0.55	0.497	411.3	477.8	8,065,680
2019	38.2	11.7	0.55	0.498	436.1	505.7	8,364,353

(b) Venezuelans with PEP

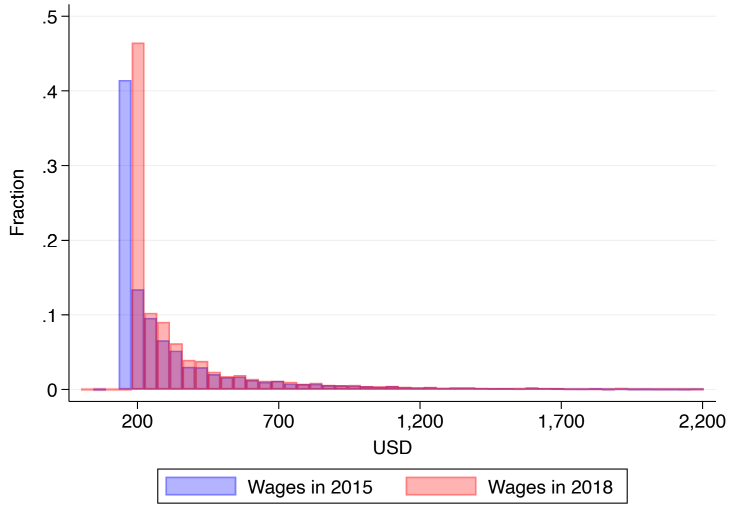
	Age		Sex (%)		Real wages (USD)		
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	N
2018	30.8	7.8	0.68	0.467	243.7	99.0	12,842
2019	31.8	8.1	0.67	0.472	248.7	98.6	42,755

Note: This table reports the descriptive statistics for Colombians and Venezuelan migrants 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 taken into account for wages and the number of observations. 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 World Bank. For self-employed workers, observed wages in PILA corresponds to 40% or more of their actual wages by law, with the minimum wage as a lower bound. Source: PILA, 2013–2019, August.

Figure 1 shows how binding is the minimum wage in Colombia for a large portion of formal workers. In 2015, around 40% of all formal workers earned the minimum wage.[10] From this figure is clear the bunching of wages for both years, 2015 and 2018. This downward rigidity helps explain why there cannot be further wage drops for these low-wage workers in the face of a supply shock.

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

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



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

3.2 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.[11] Table 2 is split into seven categories of firm size to show certain facts. First, in terms of gender, male workers are the main group in all firm sizes, especially for the 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 growing with firm size.

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

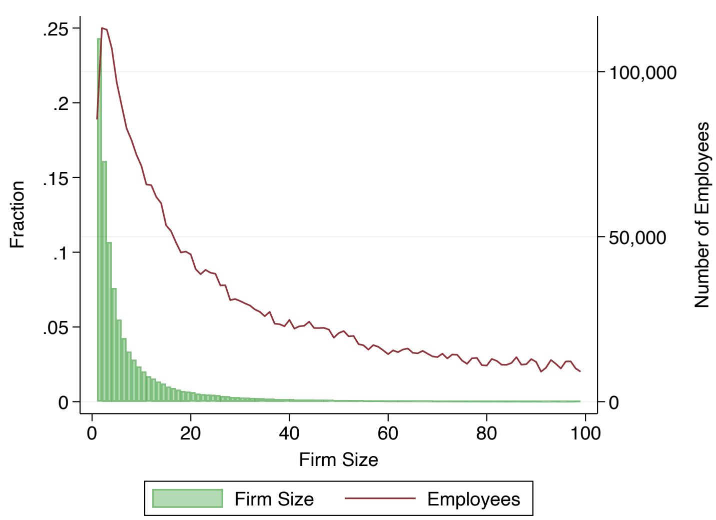
Table 2: **Descriptive statistics by firm size**

	Average				
Firm size (workers)	Employment	Sex (%)	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. MW refers to minimum wage. 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 World Bank. Only workers who contribute as employees are taken into account. Source: PILA, 2015-August.

Next, Figure 2 plots the histogram of firms by size and 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 firm sizes.

Figure 2: **Histogram of firm size and total employees**



Note: The firm size upper bound is restricted to 100 workers. Chosen binwidth is 1. Source: PILA, 2015.

4 Empirical Strategy

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

$$\frac{w_{i,l,t} - w_{i,l,2015}}{w_{i,l,2015}} = \delta_t + \theta_t \Delta M_{l,2018} + X_i' \beta + \Delta u_{i,l,t} \tag{1}$$

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 fixed effects for seven age groups with dummies for gender and self-employment.[12] Using this specification, I compare individuals that 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 below in detail. Moreover, by taking differences, individual constant unobservable characteristics are ruled out from the analysis. In this case, $\theta_{2015} = 0$ by construction.

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

$$e_{i,l,t} - \sum_{k=2013}^{2015} e_{i,l,k}/3 = \omega_t + \gamma_t \Delta M_{l,2018} + X_i' \beta + \Delta u_{i,l,t} \tag{2}$$

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 take into account the average of employment in the pre-period to weight by workers' labor trajectory in the formal sector. Note that, in the event study figures, I take the simple difference with the base period ($e_{i,l,t} - e_{i,l,2015}$) to avoid pre-treatment coefficients to be mechanically located around zero. The intercept of each year difference is given by δ_t and the standard errors in all the specifications are clustered at the level of the treatment, which are the FUAs ($G = 53$).

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

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

where the numerator is the stock of employed Venezuelans (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

[12]Industry fixed effects are not used as controls because is not possible to observe that variable in the PILA.

employed Venezuelans in l whose year of arrival was 2015. 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 downward biased (see ordinary least squares (OLS) estimates of Figure 4a and 4b). Thus, to consistently estimate the causal 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 A.1a 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, exports to Venezuela are regularly around zero for border departments. Furthermore, 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 no differential economic trends before the migration crisis started (see Figure A.1b). 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 is past settlements of Venezuelans, which is defined as:

$$z_l = \left(\frac{Ven_{l,2005}}{Ven_{2005}} * M_{2018}\right)/L_{l,2015} \tag{4}$$

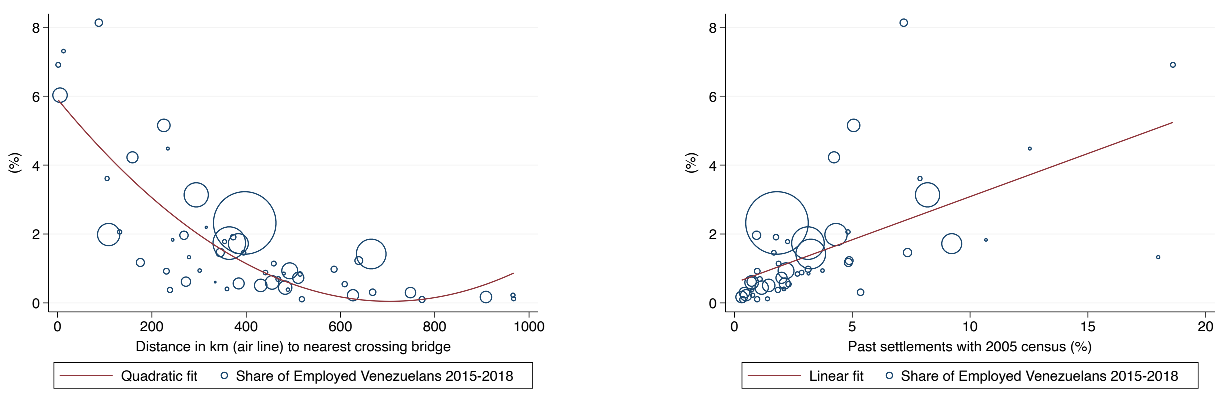
where the first term is the share of Venezuelans in FUA l (according to the 2005 population census), normalized by the formal employment $L_{l,2015}$ in l, whereas M_{2018} are the number of Venezuelans in Colombia according to the census of 2018.[13] The denominator $L_{Total,l,2018}$ is the total number of employed population in l. Past settlements are used as an instrument because newly arriving immigrants will likely move to areas with previously established Venezuelans. In order 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 3a and 3b show the first stage of the migration shock $\Delta M_{l,2018}$, for the 53 FUAs defined for this analysis, against the instruments. The instruments' relevance and functional form can be inferred from these figures. For the first instrument, a larger distance from a crossing bridge decreases the share of employed Venezuelans 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 Venezuelans that represent between 6% and 8% of their overall employed population.[14]

Figure 3: **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=53). In panel (b) four areas are removed from the graph for expositional clarity because they were outliers. Source: CNPV, 2018.

Figures 3a and 3b are constructed at the FUA level. Yet, since the aim of this paper is to estimate the impact of immigration at the individual level, the first-stage of the two-stage least

[13]The instruments is similar to the constructed in Delgado-Prieto (2022) and it is based on the canonical approach of Card (2001).

[14]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.

squares regression (TSLS) is going to weight differently each FUA by the number of individual observations available. Hence the first-stage can vary slightly depending on the sample used. With this caveat in mind, the first-stage model is:

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

In (5) two polynomials of distance to the nearest crossing bridge are included, z_l are the past settlements of Venezuelans, and the error term is υ_l (it captures the endogenous component of $\Delta M_{l,2018}$). I combine the two instruments in the analysis as past settlements or distance capture different exogenous components of migration, while increasing the F-statistic and R^2 of the first-stage regression (see Table A.1). As a result, equations (1) and (2) are estimated throughout the paper using TSLS with the aforementioned instruments.

5 Worker-level Effects

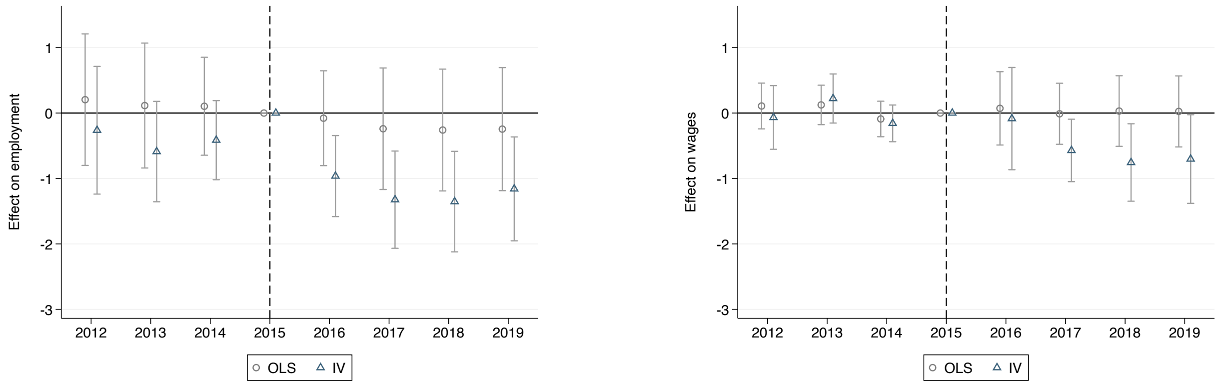
This section documents the impact of immigration on formal wages and formal employment at the worker-level. First, I show wage and employment estimates using two methods, OLS and TSLS. One advantage of the proposed empirical specification is that is possible to test for differential trends of the outcome before the immigration shock happens. Importantly, there are no significant pre-trends for employment and wages that can confound the impact of immigration.

First, 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 pp increase in the share of employed Venezuelans in a given area reduces the probability of being employed in the formal sector by –1.3 pp (see Figure 4a).[15] In terms of formal wages, I find a coefficient of –0.8% in 2018 for a one pp increase in the immigration shock (see Figure 4b).

[15] This coefficient comes from a regression with dependent variable $e_{i,l,2018} - e_{i,l,2015}$, which captures the difference in the employment indicator in 2018 relative to base period. In the heterogeneity analysis, the coefficient of employment comes from a regression using as a dependent variable the cumulative employment change $e_{i,l,t} - \sum_{k=2013}^{2015} e_{i,l,k}/3$. The second one gives slightly less negative coefficients.

Figure 4: **Event study estimates on individual wages and employment**

(a) Employment (b) Wages



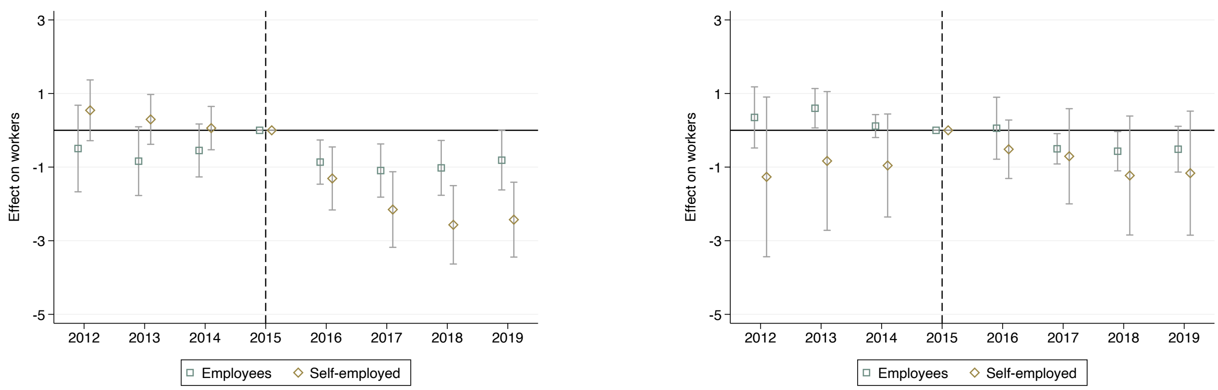
Note: The sample in panel (a) is 6,409,615 workers, while in panel (b) varies slightly year by year as the workers must be employed in the post-treatment and base year. Moreover, sample is restricted to natives between 20 and 55 years old. I use as controls interactions of gender with seven age categories and dummy of employee in the base period. I cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA 2013–2019.

For the rest of analysis, I focus on wage and employment estimates using the characteristics of workers before the immigration shock, specifically, the characteristics in 2015. So, by job type of worker there are two categories, employee or self-employed.[16] Self-employment in Colombia represents about half of the employed population, mainly working in the informal sector but with a share of workers in the formal sector that is important (around 18% of all native formal workers were self-employed in 2015). Figure 5a shows a drop in the probability of being a formal worker for self-employed natives that is more negative compared to native employees. On the other hand, the coefficient for wages of self-employed workers is quite noisy, while employees have a more precise negative coefficient (see Figure 5b).

[16] I use only IV hereafter because OLS estimates are inconsistent (see Figures 4a and 4b).

Figure 5: **Event study estimates on individual wages and employment by job type**

(a) Employment (b) Wages

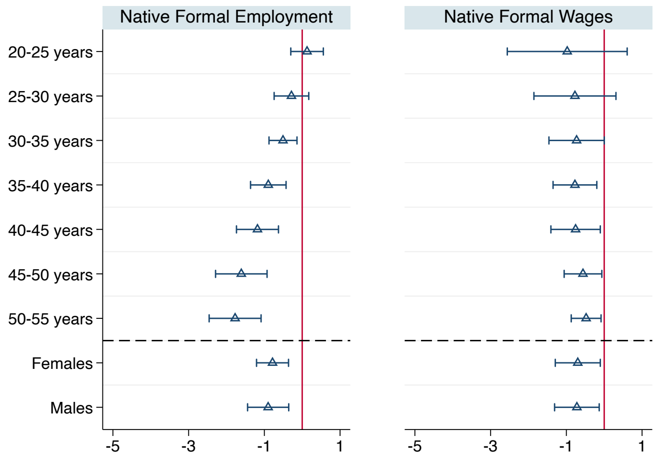


Note: The sample in panel (a) is 6,409,615 workers, while in panel (b) varies slightly year by year as the workers must be employed in the post-treatment and base year. Moreover, sample is restricted to natives between 20 and 55 years old. I use as controls interactions of gender with seven age categories and dummy of employee in the base period. I cluster standard errors (G=53). 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 the characteristics of workers in 2015 (the year before the immigration shock).[17] To begin with, the sample is divided into seven age groups and by gender, the controls used in the main specifications. Figure 6 shows a decreasing pattern in the probability of being employed in the formal sector as the worker is older. In contrast, for wages the pattern is not equally clear, and I find similar negative estimates in all age groups. Second, for gender, the impact for employment and wages is alike, there are no differential effects in this group category.

[17] Nonetheless, in Appendix Table A.4 I show there are no systematic pre-trends by age groups on employment or wages.

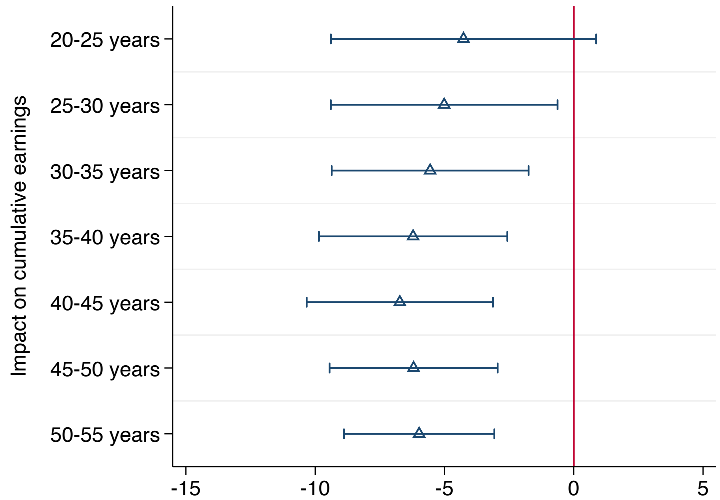
Figure 6: **Estimates by age category and gender, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

I then 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 base period. If the worker is not employed in any given period the earnings are zero, so this outcome yields a combined effect of the observed changes in employment and wages. Figure 7 shows that older workers present the most negative reduction on cumulative earnings, yet the point-estimate is not so different compared to the one of youngest workers.

Figure 7: **Impact on cumulative earnings by age group, 2015–2018**



Note: Dependent variable is $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,Pre}}$. The sample is restricted to natives between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

To complement the interesting 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 get older, their labor supply is more elastic, that is, the responsiveness to work from changes in wages is more important for older workers than for younger workers. Similar to Dustmann et al. (2017) findings for Germany, where the elasticity of labor supply is increasing in workers' age.

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

Age group	20-25	25-30	30-35	35-40	40-45	45-50	50-55
η_w^s	-.12	.36	.69	1.12	1.55	2.86	3.73

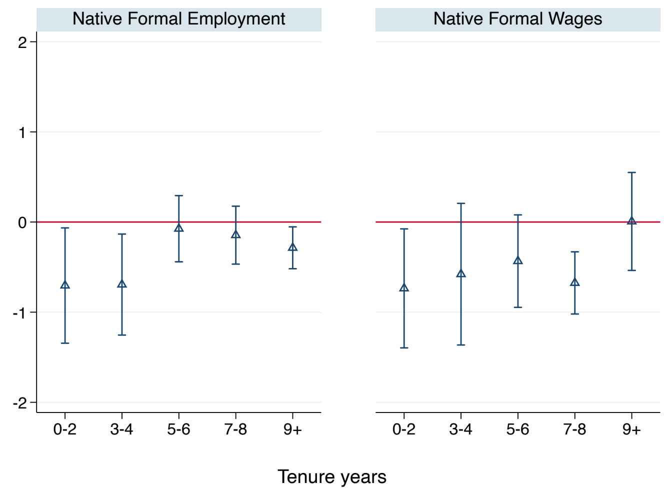
Note: The elasticity of labor supply is given by the reduced-form results from changes in native employment over changes in native wages.

To understand which are the most affected workers, I exploit the number of years the worker has been employed in the same firm (i.e., job tenure) up to the base period of 2015.[18] Figure 8 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

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

with fewer years in the same firm. This pattern is more precise on employment than on wages. This result can be partly explained by the accumulation of firm-specific human capital of native workers, as they can be less substitutable to migrants with similar characteristics.

Figure 8: **Estimates by job tenure, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives employees between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

The last two figures suggest that older workers and workers with lower tenure have the most significant drop on formal employment from the immigration shock. To understand with more detail which are the most affected groups I now combine the worker's age with their job tenure. Table 4 shows that native workers below 35 years old present an insignificant effect on employment, independent if they have low or high job tenure. On the other hand, native workers above 35 years old present a significant negative effect on employment when they have low and high job tenure, and the effect is much higher for the workers with lower tenure (–1.2 pp versus –0.36 pp). In terms of wages, only younger worker with high tenure present a significant negative effect, more concretely, a 1 pp increase in the immigration shock reduces their wages by 0.86%.

Table 4: **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	-0.144	0.254	-1.200**	-0.360**
	(0.230)	(0.296)	(0.384)	(0.117)
N	2,010,695	330,636	1,977,880	1,040,248
Wages	-0.628	-0.862*	-0.723	-0.135
	(0.402)	(0.357)	(0.430)	(0.182)
N	1,049,714	230,620	1,116,544	755,077
Clusters	53	53	53	53

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

Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives employees between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2018.

5.1 Mobility Responses

Another benefit of individual data compared to regional data, is the possibility to estimate inter-regional movements of each worker as a response to the immigration shock. Table 5 shows the probability that an employed native worker changed region in 2018 relative to 2015. Results suggest that formal workers are not moving to other regions after migrants arrive (both coefficients are not significant).[19]

Table 5: **Estimates on regional changes of formal workers, 2015–2018**

	OLS	IV
Prob. of changing region	-0.204	-0.033
	(0.322)	(0.359)
N	3,914,830	3,914,830
Clusters	53	53

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 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. The PILA has a measurement error with the location variable in 2018, that is not related to the instruments, still the estimates are interpreted as suggestive. Workers are observed in August of each year. Source: PILA, 2015–2018.

Next, as discussed in Dustmann et al. (2017) younger workers tend to be more mobile. Hence,

[19] Hence, the estimates at the regional-level for the formal sector in Delgado-Prieto (2022) are not attenuated by the movement of natives across regions.

Table 6 shows the impact of movements across regions by age groups. Indeed, younger formal workers tend to move more from higher exposed areas to immigration, but the coefficients are not significant. Overall, the point-estimates decrease as the worker gets older, but all of them are insignificant. The mobility margin of adjustment is less of a concern in this setup.

Table 6: **IV estimates on regional changes of formal workers by age group, 2015–2018**

Age group	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Prob. of changing region	0.231	0.088	0.061	-0.024	-0.143	-0.183	-0.192
	(0.436)	(0.418)	(0.425)	(0.388)	(0.326)	(0.307)	(0.257)
N	138,110	750,744	773,840	668,208	573,570	543,266	467,198
Clusters	53	53	53	53	53	53	53

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 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. The PILA has a measurement error with the location variable in 2018, that is not related to the instruments, still the estimates are interpreted as suggestive. Workers are observed in August of each year. Source: PILA, 2015–2018.

5.2 Worker-level and Regional-level Estimates

Following the framework of Dustmann et al. (2022b), in here I compare worker-level estimates with regional-level estimates of Delgado-Prieto (2022). In terms of employment, is possible to disentangle the regional formal employment estimate into a displacement of incumbent workers (outflows from formal employment), hirings of new workers (inflows to formal employment) and reallocation of existing employed workers. The first distinction is that worker-level estimates of employment capture only the outflows or the displacement of incumbent native workers. For that reason, the coefficient of formal regional employment in Delgado-Prieto (2022) is –2.2 and the coefficient on individual formal employment is –1.3.

In terms of wage estimates, the worker-level response is negative and significant, while the regional-level estimate in Delgado-Prieto (2022) is insignificant and close to zero. These two responses are complementary and answer to different policy questions, in any case the differential estimate can be rationalized as follows. The immigration shock can change the composition of employed natives, positively selecting the individuals that remain in the region and therefore increasing regional wages. On the other hand, immigration can decrease the price of labor reducing regional wages. In the worker-level regressions, the wage estimate only captures the change in the price of labor holding the composition of the population constant (Dustmann et al., 2022b). On the opposite, the regional-level regressions measures jointly the change in selection of workers and change in prices of labor. Hence, it is possible to find 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.

5.3 Distributional Impacts of Immigration

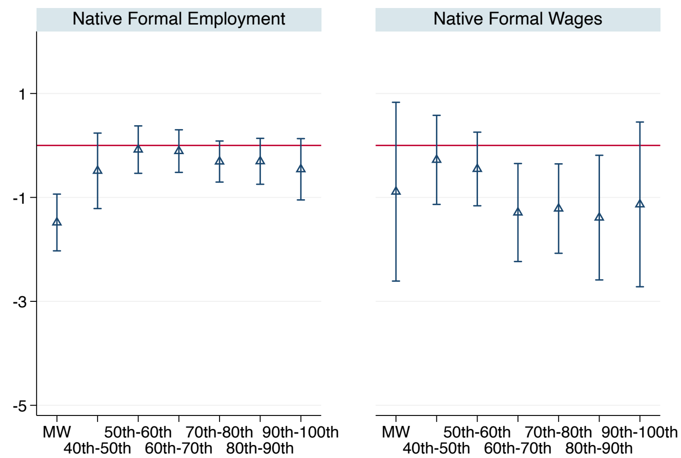
I then estimate the impact of immigration on workers across the distribution of wages. All native workers employed in 2015 are divided according to their local wage distribution in that year. Figure 9 plots the coefficients for 2018 relative to 2015 according to this division.[20] It is explicit the uneven impact of immigration: native workers earning the minimum wage suffer the highest 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 Venezuelans in a given labor market reduces the probability of being employed in the formal sector by −1.3 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 binds, there cannot be further drops of formal wages, and this muted wage effect augments the employment effect. Of course, movements to the informal sector of these low-wage workers, which cannot be observed, explain part of the large coefficient.

Another fact that is muted in a regional-level analysis of Delgado-Prieto (2022) is the possibility of relative wage losses in the formal sector. For the workers between the 60th and 90th percentile of the local wage distribution, I find a drop of around −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. Last, the exit of low-wage workers from formal employment accompanied by an average drop of wages from workers in the middle and upper part of the wage distribution can explain the insignificant effect

[20] As a robustness check, Appendix Table A.5 shows the pre-treatment coefficients by wage categories on employment and wages. Reassuringly, most of these coefficients are insignificant.

on regional formal wages found in Delgado-Prieto (2022).

Figure 9: **Estimates by wage category, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. Wage distribution categories at base period. The sample is restricted to natives between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

6 Worker-level Effects by Firm Characteristics

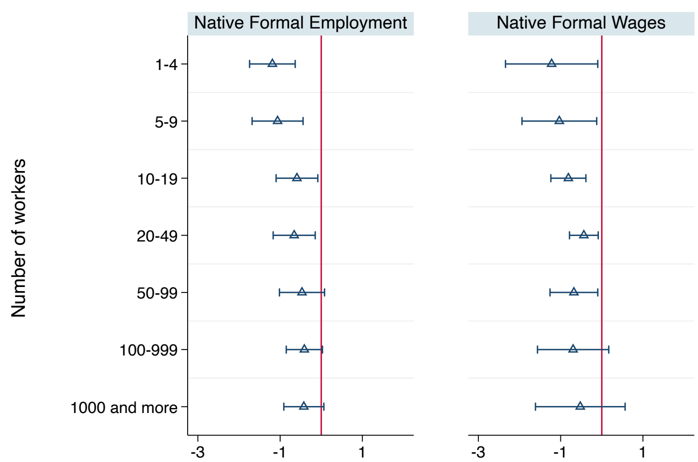
After analyzing the effects by individual characteristics, I turn to the firm information for each worker to provide complementary results of the channels of adjustments to immigration. To begin with, Delgado-Prieto (2022) shows that Venezuelan immigrants are disproportionally employed in the smallest firms.[21] So, 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 10 shows that workers in firms with less than 50 workers present a significant negative effect on the probability of being a formal worker, while workers in bigger firms do not have a significant negative effect. To explain this finding, besides migrants' working in the smallest firms, it is also because Colombia has a dual-labor market with formal and informal employment. A key difference is that informal employment has flexible wages, while formal employment has a binding minimum wage. In Delgado-Prieto (2022), there is compelling evidence that the labor supply shock

[21]Migrants can be either formal or informal and can be self-employed.

of Venezuelan immigrants reduced significantly informal wages. But, most importantly, it is shown that smaller firms substitute formal for informal labor when both workers are strong substitutes in response to lower informal wages. Because the cost of being caught by authorities is low for the smallest firms compared to the largest firms, it is expected that native workers in the smallest firms have a more negative employment shock compared to native workers in bigger firms. Next, for wages, native workers in firms with less than ten workers present the most negative coefficient, yet all workers in firms with less than 100 workers present a significant negative effect.

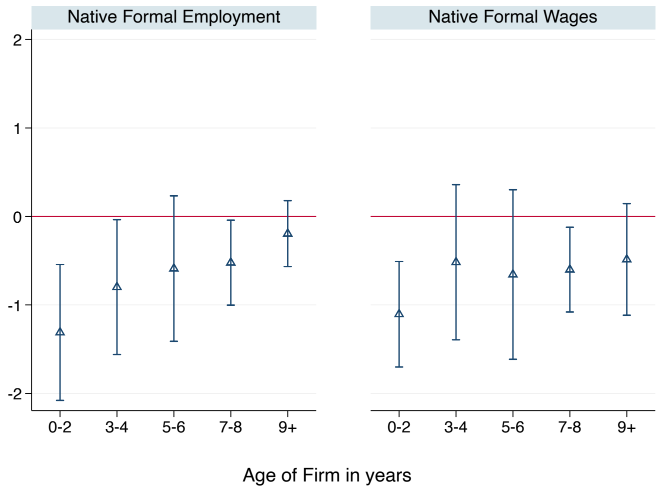
Figure 10: **Estimates by firm size, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Workers are observed in August of each year. Source: PILA, 2013–2019.

Next, the number of years the firm appears in the administrative records up to the base period, a proxy of firms' age, is used for the analysis. Figure 11 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 so striking for wages. Still, workers in the youngest firms present the most negative coefficient on wages.

Figure 11: **Estimates by age of firm, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. Sample is restricted to natives employees between 20 and 55 years old. The age of firm is the number of years the firm appears discontinuously in PILA. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

The positive correlation between firm's size and age of firm can help to explain previous negative findings, as smaller firms tend to be younger. For that reason, I combine those characteristics to determine the differential impact of immigration on smaller and larger firms by their age. Table 7 shows that native workers in the smaller firms present a significant negative effect on employment and wages, independent if their firm is young or old, but the coefficient for wages is more negative in younger firms. On the other hand, native workers in the larger firms present a significant negative effect on employment and wages only in the younger firms.

Table 7: **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	-0.938**	-0.941***	-1.204**	-0.187
	(0.358)	(0.222)	(0.421)	(0.208)
N	454,328	475,043	876,294	3,553,794
Wages	-1.412**	-0.754*	-0.680*	-0.523
	(0.533)	(0.349)	(0.334)	(0.357)
N	260,746	335,627	421,488	2,134,094
Clusters	53	53	53	53

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

Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives employees between 20 and 55 years old. The age of firm is the number of years the firm appears discontinuously in PILA Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2018.

6.1 AKM Model

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. To construct them is needed to estimate the standard AKM framework proposed in Abowd et al. (1999). This model decomposes the contribution of firm-specific and worker-specific constant characteristics to log formal wages (lnw_{it}). The AKM model is expressed as

$$lnw_{it} = \alpha_i + \psi_j(i,t) + X_{it}'\beta + \epsilon_{it} \tag{6}$$

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 three polynomials of age, year fixed effects and a dummy for the capital, and 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 fixed effects that workers move across firms in the period of analysis and, for consistency, it is needed a 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 effects $\hat{\psi}_1, ..., \hat{\psi}_J$ and worker effects $\hat{\alpha}_1, ..., \hat{\alpha}_N$. Note that Bonhomme et al. (2020) and Card et al. (2018) have documented a limited mobility bias in AKM models, especially for smaller firms. This limitation can exist in this model, and one solution would be to eliminate all the small firms from the estimation.[22] However, as migrants are more likely to be employed in the smallest firms, I prefer to focus on all firms, possibly having a mobility bias.[23]

To begin with, Appendix Table A.7 shows the decomposition of the variance of wages $Var(lnw_{it})$ in the formal sector of Colombia. Worker effects explain 78% of the variance and firm effects explain 16%, 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, this sorting explains an additional 11% of the variance.

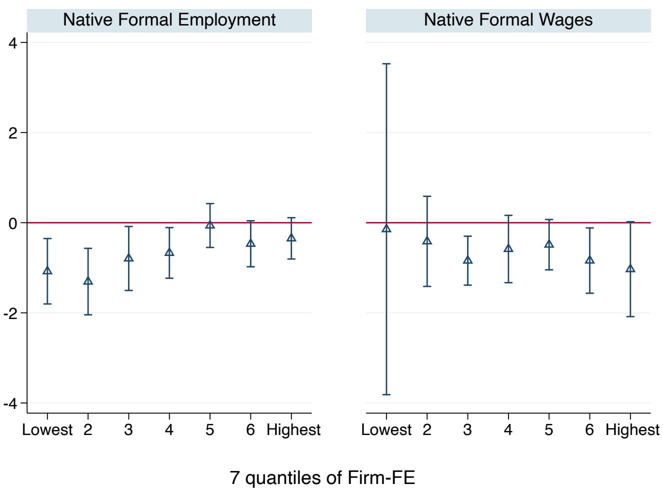
Then, using the estimated $\hat{\psi}_j$, I divide workers by seven quantiles of firm fixed effects or firm-specific pay premiums, which I now refer to as lowest- or highest-paying firms, to compute the impact of immigration.[24] Figure A.4 shows that workers at lowest-paying firms in 2015 suffer the most negative employment losses while having insignificant changes in wages. In contrast to workers in highest-paying firms in 2015, where the wage response is more negative than the employment response, similar to the findings in Gulyas et al. (2019) for a mass layoff shock in Austria. A possible explanation for this result is that there are lower outside options for natives in high-paying firms after migrants arrive. This strengthens the bargaining power of firms and might incentivize lower wage growth for its workers. Last, to define which adjustment, if wages or employment, can decrease more overall earnings, I use the cumulative earnings outcome defined previously. Figure 13 shows that workers in the lowest and highest paying firms are similarly affected in terms of earnings, either via wage or employment losses, while workers in the middle of the pay premiums are the least affected.

[22] Another solution is to aggregate small firms according to their observable characteristics, but since I do not have information on the industries they operate, the aggregation will include high productivity sectors with low productivity ones, misleading the estimates.

[23] 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 effects with their corresponding covariance) but does not estimate a corrected vector of $\hat{\psi}_j$, which is what is used in this paper.

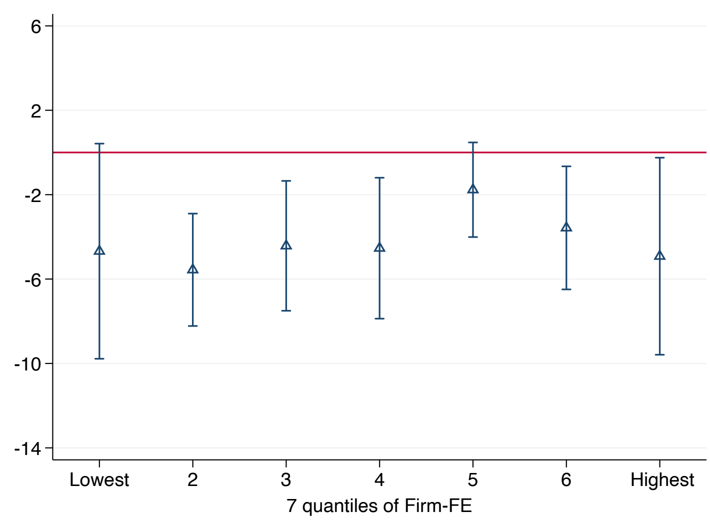
[24] The normalization process of the $\hat{\psi}_j$ states that the average of firm effects is equal to zero across groups. The groups are defined as the combination of workers who have been employed in certain firms with their respective firms. Workers who are not employed by these previous firms belong to other groups, along with the firms that employed these workers. So, to compare $\hat{\psi}_j$ properly I need to assume that the average firm effect in each group is the same.

Figure 12: **Estimates by firm quantiles of fixed effects, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. Sample is restricted to natives employees between 20 and 55 years old. Firm fixed effects are computed in a first-step using the standard AKM framework, with age and its cubic as time-varying controls, for the period 2013-2015. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. 95% confidence interval. Source: PILA, 2013–2019.

Figure 13: **Estimates for cumulative earnings by firm quantiles of fixed effects, 2015–2018**

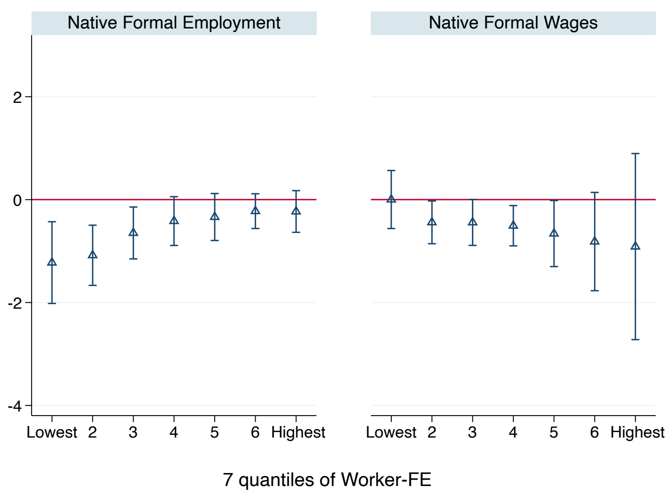


Note: Dependent variable is $\sum_{t=2016}^{t=2018} \frac{Earnings_{it}}{Earnings_{i,Pre}}$. Sample is restricted to natives employees between 20 and 55 years old. Firm fixed effects are computed in a first-step using the standard AKM framework, with age and its cubic as time-varying controls, for the period 2013-2015. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. 95% confidence interval. Source: PILA, 2013–2019.

Next, Figure 14 shows the same exercise but dividing by seven quantiles of worker fixed effects:

$\hat{\alpha}_i$. The wage and employment estimates hold similarly as before. High-quality workers present the most negative point estimate for wages and the least negative one for employment, in contrast to lower-quality workers where the wage effect is close to zero, and the employment effect is more negative. It is clear how wages and employment are mirror images of each other and should be studied together as emphasized in Dustmann et al. (2017).

Figure 14: **Estimates by worker quantiles of fixed effects, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. Sample is restricted to natives employees between 20 and 55 years old that appear more than once in PILA. Worker fixed effects are computed in a first-step using the standard AKM framework, with age and its cubic as time-varying controls, for the period 2013-2015. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. 95% confidence interval. Source: PILA, 2013–2019.

6.2 Interaction of Worker and Firm Characteristics

As shown previously, workers' and firms' characteristics before the immigration shock determine a differential impact of immigration on wages and employment. In Table 8, I combine the previous responses by restricting the sample to the subgroups where previous findings indicate a more negative coefficient on employment. First, for minimum wage earners in 2015, immigration reduces the probability of employment in the formal sector by 1.7 pp. For the medium age group, the impact is less negative (–1.4 pp), while for self-employed workers, the impact is more negative (–2.7 pp). When combining these three characteristics, there are 535,969 workers in the sample, for whom the effect of Venezuelan immigration on the probability of being a formal worker is of -3.3

pp, a large displacement effect.

Table 8: **Most affected native workers in terms of employment, 2015–2018**

	(1)	(2)	(3)	(4)	(5)
Prob. of employment	-1.019***	-1.750***	-1.447***	-2.687***	-3.228***
	(0.250)	(0.304)	(0.297)	(0.433)	(0.491)
Sample restriction					
Minimum wage earners	✗	✓	✗	✗	✓
Medium age (35 years or more)	✗	✗	✓	✗	✓
Self-employed	✗	✗	✗	✓	✓
N	6,409,615	2,109,035	3,737,922	1,050,156	535,969
Clusters	53	53	53	53	53

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 are the coefficients, 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 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2013–2019.

Again, to combine previous responses, I use the same criteria as in Table 9 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 0.8%. For workers in the smallest firms in 2015, the impact is more negative (–1.1%), while for workers in high-paying firms in 2015, I find an estimate of -0.84%. When combining the first two characteristics, there are 281,893 workers in the sample, for whom the effect on wages in 2018 is –2% for a one pp increase in the immigration shock.

Table 9: **Most affected native workers in terms of wages, 2015–2018**

	(1)	(2)	(3)	(4)	(5)
Wages	-0.756*	-0.949*	-1.101**	-0.840*	-2.088***
	(0.302)	(0.395)	(0.376)	(0.365)	(0.469)
Sample restriction					
Above minimum wage	✗	✓	✗	✗	✓
Small firm (1 and 19 workers)	✗	✗	✓	✗	✓
High-paying firm (quantile 6)	✗	✗	✗	✓	✗
N	3,914,830	2,537,836	611,665	95,473	281,893
Clusters	53	53	53	53	53

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 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2015–2018.

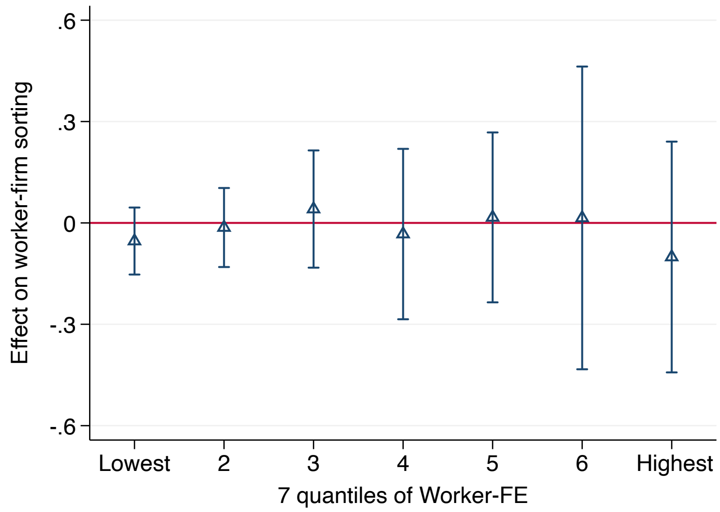
6.3 Sorting

Last, I study the reallocation effects from the immigration shock analyzing changes in the sorting patterns of high- and low-paying workers into high- and low-paying firms.[25] In this exercise, the outcome is constructed using the values of $\hat{\psi}_j$ from equation (6) and exploiting the movements of workers between firms in the post-treatment period. More concretely, the outcome is the change in the AKM firm fixed effects 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.[26] Results are shown by seven quantiles of worker fixed effects 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 15 plots the estimates for these categories, and none of them present significant results. There are no reallocation effects of immigration. 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., 2022a).

[25]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.

[26]Since the fixed effects are constructed for the pre-policy period, all workers in firms created after 2015 are not taken into account in the analysis. The estimated firm effects are turn to positive values for the construction of the outcome.

Figure 15: **Estimates by worker quantiles of fixed effects, 2015–2018**



Note: Dependent variables are cumulative employment and wages relative to the base period. The sample is restricted to natives between 20 and 55 years old. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Workers are observed in August of each year. Source: PILA, 2013–2019.

7 Machine-Learning Approach

In this section of the article, a machine learning algorithm is proposed to determine the differential role of workers and firms in the impact of immigration. In the last section, wage and employment effects are shown for specific 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 et al. (2019) and implemented by Gulyas et al. (2019). This framework can identify the subgroups who experience the greatest wage and employment losses by a recursive partitioning method. 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. I use the `grf` package in `R` to estimate the causal forests. The specification proposed combines DiD with IV, so the regression estimated is:

$$\Delta Y_{i,l,2018} = \tau \hat{M}_{l,2018} + \Delta \epsilon_{i,l,2018} \tag{7}$$

where ΔY is the outcome of interest, the difference of individual wages or employment in 2018 relative to 2015, matching the census rate $\hat{M}_{l,2018}$, and this is already the predicted immigration rate after regressing the observed one with the two instruments. A vector 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, gender, job tenure, wages, firm effects, worker effects, and firm size.[27]

The procedure in Athey et al. (2019) consists of several steps that are adapted to this setup. The algorithm proceeds as follows:

1. Start with 10% of the full sample due to computational burden, name this full sample P.

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 each of the variables in X_f, the data is split in two partitions (say P_l and P_r) to run separate regressions of (7) to estimate treatment effects for each partition. Choose the variables with their threshold values that maximize the difference in treatment effects using this formula:
$$(\tau_l - \tau_r)^2 \frac{n_l n_r}{N^2} \tag{8}$$
where n_l and n_r refers to sample size of each partition and full sample refers to N.

4. Recursively form the resulting nodes to the left and to the right of the decision tree until the nodes reach a minimum node size, the difference in sample size is too big, or when the split would only yield a difference in treatment effects relatively small.

In this procedure, the variables that appear more frequently as splits in the forest are defined as more important to explain treatment effect heterogeneity. I perform the algorithm excluding and including firms' variables to show the difference in the importance measure.[28] First, 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 16a). However, when

[27]The procedure sample varies depending on the features selected but starts with the same sample always. 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 going to be smaller.

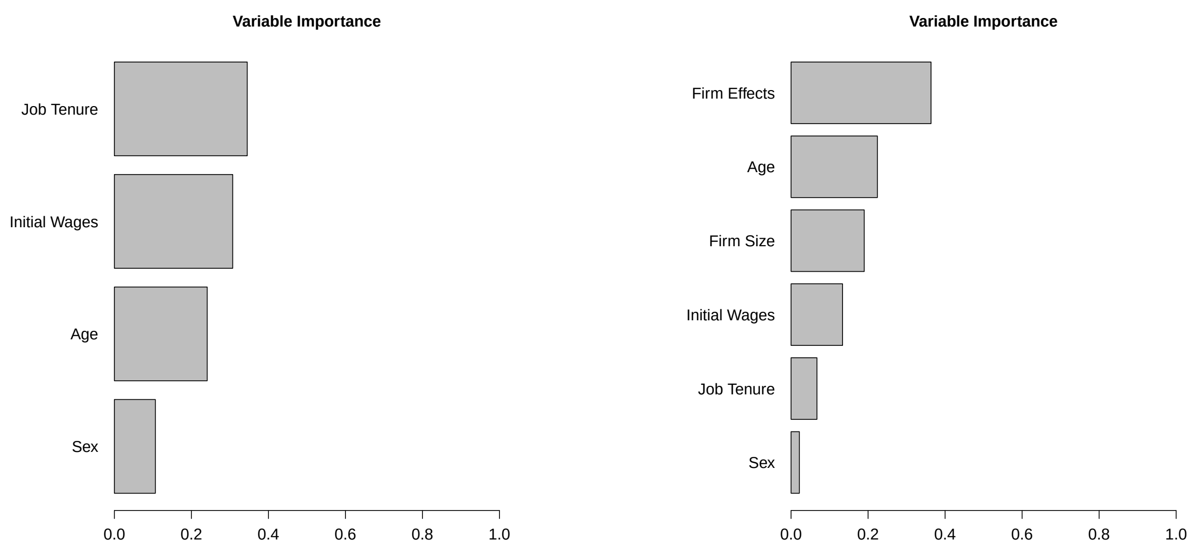
[28]For employment, I use the individual change in employment between 2018 and the average pre-period employment as outcome.

including firms' variables, the most important variable becomes firm-specific pay premiums or firm fixed effects followed by firm size and initial wages (see Figure 16b).

Figure 16: **Variable importance for employment in causal forest, 2015–2018**

(a) Without Firm Variables

(b) With Firm Variables



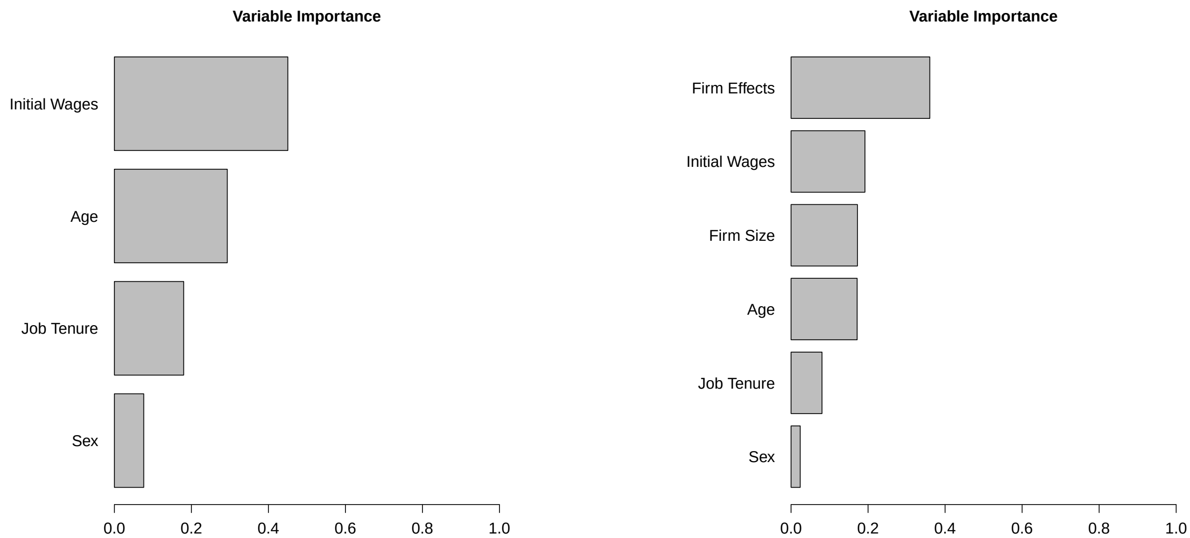
Note: Variable importance is a weighted sum of how many times the feature f appears in split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 20 and 55 years old. All values sum up to 1. I use clusters of FUA for the causal forest estimation. The minimum node size is 300.

Following up, I construct the individual percent change in wages between 2018 and 2015, to perform the same exercise as before. Without firms' variables, the most important variables are initial wages followed by age (see Figure 17a). However, when including firms' variables in the causal forest, again, firm-specific pay premiums followed by age and initial wages are the most important variables (see Figure A.2b).

To summarize, the most important variable to explain wage and employment changes relates to the firm-specific pay premiums or firm fixed effects more than any worker characteristics. In the causal forest of wages, firm effects appear in 31% of all splits, for employment, firm effects appear in 29% of the splits. Note that, with this algorithm, I do not recover any average partial effect. Its main use is to determine the relative importance of all the variables analyzed in the reduced-form estimates.

Figure 17: **Variable importance for wages in causal forest, 2015–2018**

(a) Without Firm Variables (b) With Firm Variables

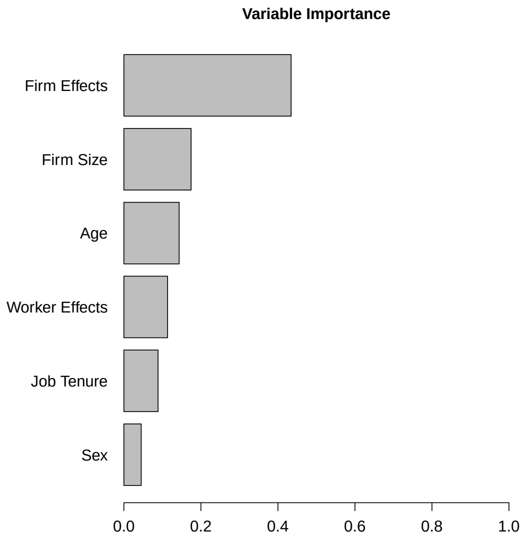


Note: Variable importance is a weighted sum of how many times the feature f appears in split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 20 and 55 years old. All values sum up to 1. I use clusters of FUA for the causal forest estimation. The minimum node size is 300.

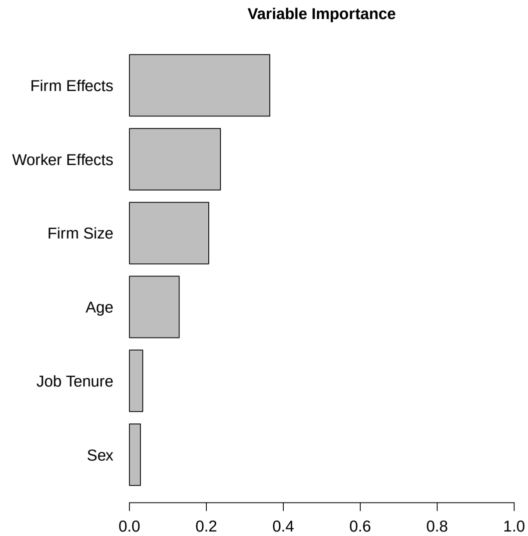
Last, as initial wages are a function of the unobserved firm and worker effects, the next step is to include in the algorithm the constructed worker effects $\hat{\alpha}_i$ instead of initial wages. This reduces the sample as every worker must be observed more than once. After adding worker effects, again the firm-specific pay premiums or firm effects are the most important variable to explain the heterogeneity of treatment effects for employment and wages (see Figures 18a and 18b). Reassuringly, firms' role in the impact of immigration is critical.

Figure 18: **Variable importance for employment and wages in causal forest with worker and firm effects, 2015–2018**

(a) Employment



(b) Wages



Note: Variable importance is a weighted sum of how many times the feature f appears in split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 20 and 55 years old. All values sum up to 1. I use clusters for the causal forest estimation. The minimum node size is 300.

8 Robustness Checks

The exclusion restriction of the distance instrument establishes a unique channel it can affect wages and employment and is through immigration. 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. For that reason, all the border areas are removed from the estimation sample and I find similar point estimates but not significant for wages (see Appendix Table A.2, row 2). Next, another concern is the relevance of Bogotá as the capital of Colombia (the proportion of observations from the capital is 32.6%) 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 A.2, row 3).

Next, further controls are added to the estimation to have a more precise comparison of 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 slightly less negative for employment, partly because for self-employed workers there is no measure of job tenure and thus they are excluded. The next robustness test relates to adjusting nominal wages to real terms using the national CPI. In this case, the results of wages are a bit less negative. Last, since the administrative wages from PILA are not censored or bounded in the right tail, I top code wages manually after the 99% percentile of the wage distribution and find that estimates are unaltered.

Third, there are robustness checks for the machine learning algorithm. In the causal forests, the number of possible values that a variable can take might alter the variable importance weighted sum. For instance, when variables take a small set of values they might mechanically appear in fewer nodes further in the tree. For that reason, Appendix Figure A.2a shows that when transforming all the continuous variables into seven categories, the order of importance changes slightly, with firm-specific pay premiums being second for wages and employment. Two other critiques of this approach are worth mentioning. The first one is that the frequency of splits in the first nodes of the tree are weighted the same as the frequency of splits in the last nodes of the tree, where the sample size is much smaller. This critique can be alleviated by using a decay exponent in the variable importance that puts more weight to the splits selected first.[29] After computing the variable importance, the order is fairly similar for wages, but for employment, the first and second variable are flipped. Firm size is the most important variable followed by firm effects, still both variables capture the role of firms (see Appendix Figures A.3a and A.3a). The next critique is that the tree is built to maximize the squared difference in treatment effects but without analyzing if pre-trends are either significant or not for all these subgroups. Probably when the treatment effect is higher, there can be differing pre-trends, however, as the algorithm is constructed does not allow to correct or check for pre-trends in every subgroup.

9 Conclusion

This is the first paper that exploits the labor supply shock of Venezuelan immigrants with the universe of formal workers in Colombia. This is an advantage in several dimensions:

1. By using an administrative panel data, is possible to measure worker-level adjustments that

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

are different to the most usual regional-level analysis of migration.

2. This allows to estimate a rich set of heterogeneous effects by the worker and firm characteristics that help to explain the mechanisms behind labor supply shocks.

3. With the help of a machine learning algorithm is possible to determine the differential role of workers and firms in the impact of immigration.

Overall, findings suggest that native workers suffer a negative employment shock from the Venezuelan immigration. However, this coefficient masks many heterogeneous responses. Specifically, low-wage workers earning the minimum wage are crowded out of the formal sector, while workers above in the wage distribution are not affected. Also, self-employed workers earning the minimum wage and above 35 years old have the most negative impact. Complementary, there is a reduction, that is not as robust or sizable as the employment response, for wages. Still, I find that workers in high-paying firms present a negative wage adjustment, while workers in low-paying firms do not experience a reduction in wages. To uncover the mechanisms behind these estimates, a causal forest is used to determine which variable is most important to explain the heterogeneity in treatment effects. Throughout this analysis, firm-specific pay premiums or firm effects appear prominently as the most important variable. In conclusion, focusing only on workers' observables when analyzing the labor market impacts of immigration might lead to an incomplete view of the primary sources of adjustments to immigration.

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.

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

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.

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.

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

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.

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.

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.

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

Dustmann, C., Otten, S., Schönberg, U., and Stuhler, J. (2022b). Labor Market Effects of Immigration ? 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.

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.

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*.

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*.

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

Peñaloza Pacheco, L. (2019). Living with the neighbors: the effect of venezuelan forced migration on wages in colombia. Technical report, Documento de Trabajo.

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

Santamaria, J. (2020). ?when a stranger shall sojourn with thee?: The impact of the venezuelan exodus on colombian labor markets.

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.

A Appendix

A. First stage

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

	(1) $\Delta M_{l,2018}$	(2) $\Delta M_{l,2018}$	(3) $\Delta M_{l,2018}$
Distance (/100)	-0.886		-0.405
	0.471		0.508
Distance (/100) squared	0.048		0.006
	0.043		0.045
Past settlements		0.104***	0.063*
		0.015	0.029
Constant	4.698***	1.612***	3.298*
	1.185	0.270	1.366
R^2	0.458	0.349	0.531
F	8.945	50.47	60.83
N	53	53	53

Standard errors below point estimates. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

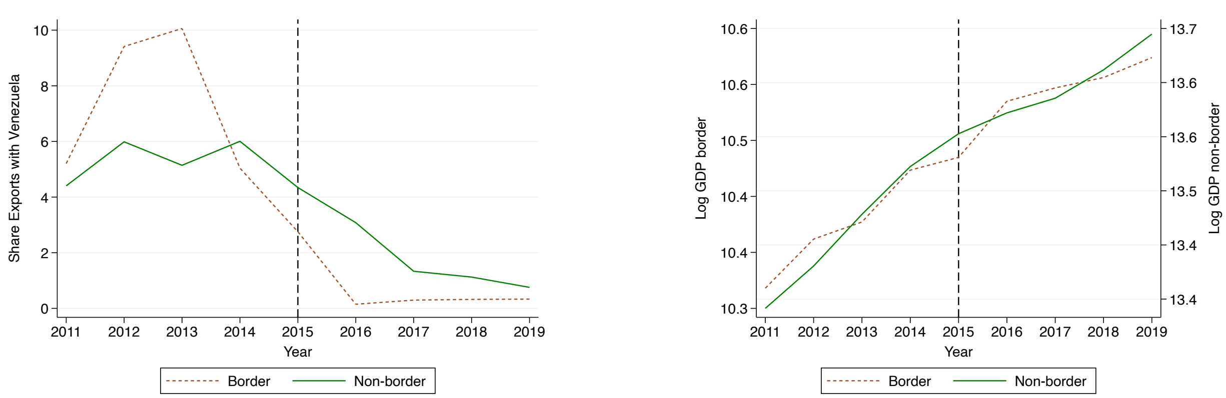
Note: This table reports the coefficient of the first-stage of the share of employed Venezuelans $\Delta M_{l,2018} * 100$ with distance and distance squared to the nearest crossing bridge and past settlements as explanatory variables. In column (1) and (3) a joint F-statistic for distance and its squared being equal to zero gives a p-value smaller than 0.001. Functional Urban Areas in Colombia (G=53) are weighted by the total number of formal workers in each area according to PILA in 2015.

B. Robustness Checks

Figure A.1: **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 A.2: **Robustness checks for wages and employment, 2015–2018**

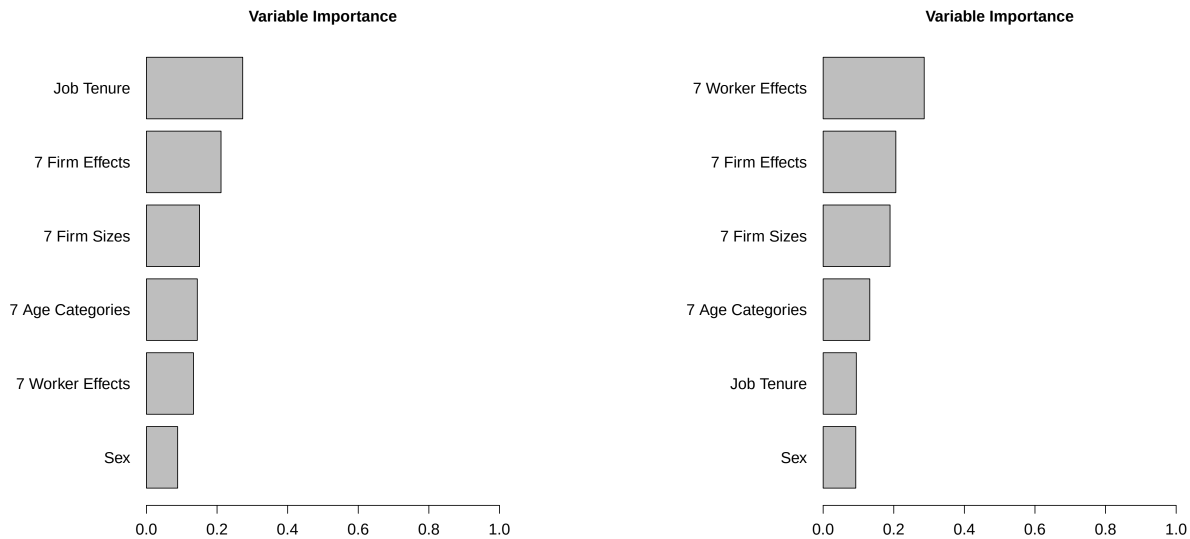
	Employment	Wages
Baseline	-1.019***	-0.756*
	(0.250)	(0.302)
N	6,409,615	3,914,830
Removing border areas*	-1.141*	-0.835
	(0.499)	(0.605)
N	6,281,503	3,839,505
Removing Bogotá	-0.922***	-0.567**
	(0.214)	(0.193)
N	4,082,309	2,465,831
Further controls$\star$	-0.530*	-0.679*
	(0.216)	(0.332)
N	4,781,526	3,151,955
Real wages		-0.638*
		(0.254)
N		3,914,830
Top code local wages above 99%		-0.742*
		(0.296)
N		3,914,830
Clusters	53	53

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 gender with seven age categories and dummy of employee in the base period. *The border areas are Cucutá, Maicao and Arauca. $\star$Further controls refer to fixed effects of seven wage quantiles and job tenure, omitting self-employed workers. The sample is restricted to natives between 20 and 55 years old. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2015–2018.

Figure A.2: **Variable importance for employment and wage losses in causal forest for categories, 2015–2018**

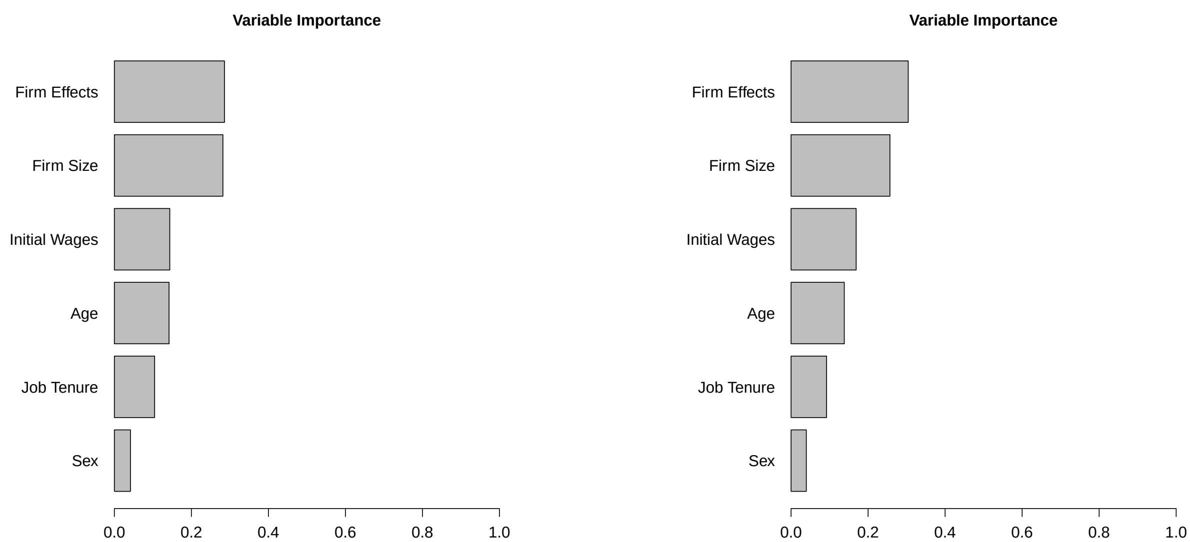
(a) Employment (b) Wages



Note: Variable importance is a weighted sum of how many times the feature f appears in split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 20 and 55 years old. All values sum up to 1. I use clusters for the causal forest estimation. The minimum node size is 300.

Figure A.3: **Variable importance for employment and wage losses in 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 split of each leaf of every tree in the forest. The number of trees is 2,000. The sample is restricted to natives between 20 and 55 years old. All values sum up to 1. The decay exponent is -2. I use clusters for the causal forest estimation. The minimum node size is 300.

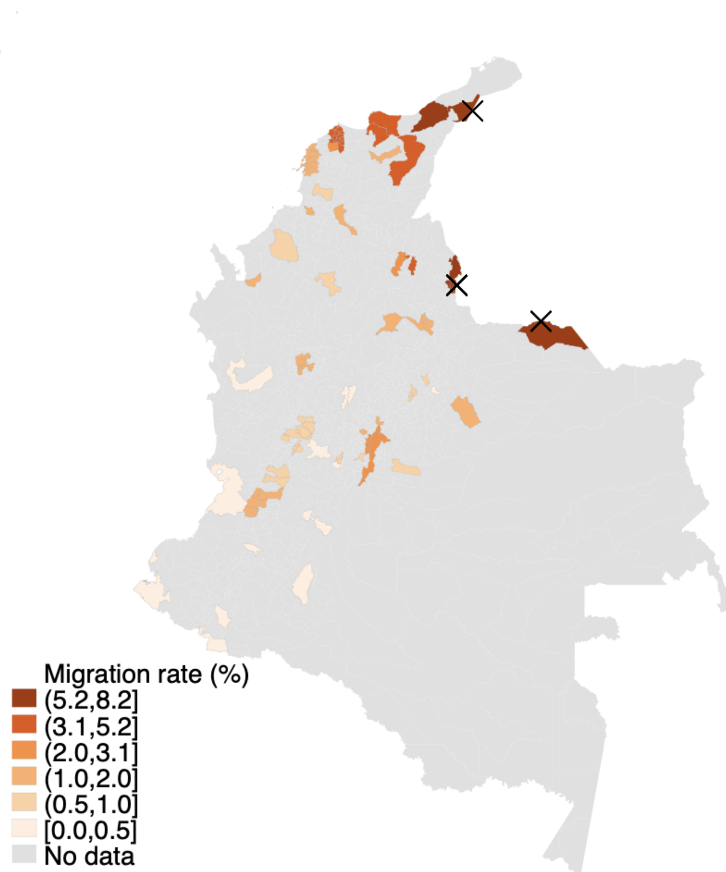
C. FUAs

Table A.3: Number of observations by FUA

	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)
No FUA assigned	713,577	(10.0)
Total	7,123,192	(100)

Note: This Table reports the number of workers from PILA by FUAs. 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 20 and 55 years old. Workers are observed in August of each year. Source: PILA, 2015.

Figure A.4: **Map of FUAs with the immigration shock**



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

D. Testing for Pre-Trends

This subsection of Appendix tests for differential trends on the outcomes according to different worker's characteristics.

Table A.4: Event study estimates on pre-treatment periods of Figure 6

	Employment			Wages		
	2012	2013	2014	2012	2013	2014
20 to 25 years	-0.596	-0.658	-0.477	-0.456	-0.080	-0.051
	(0.863)	(0.839)	(0.642)	(0.717)	(0.545)	(0.225)
25 to 30 years	-0.470	-0.822	-0.379	0.510	0.792***	0.163
	(0.747)	(0.639)	(0.484)	(0.401)	(0.228)	(0.136)
30 to 35 years	-0.643	-0.923	-0.397	-0.229	0.048	-0.016
	(0.649)	(0.508)	(0.363)	(0.356)	(0.234)	(0.199)
35 to 40 years	-0.324	-0.672	-0.531	-0.236	0.089	-0.345
	(0.524)	(0.420)	(0.351)	(0.389)	(0.390)	(0.240)
40 to 45 years	-0.036	-0.370	-0.509*	-0.054	0.262	-0.477
	(0.371)	(0.281)	(0.225)	(0.412)	(0.372)	(0.277)
45 to 50 years	0.114	-0.107	-0.153	-0.507	0.047	-0.158
	(0.337)	(0.243)	(0.194)	(0.453)	(0.383)	(0.267)
50 to 55 years	0.613	-0.037	-0.040	1.176*	1.373***	-0.177
	(0.319)	(0.194)	(0.148)	(0.463)	(0.414)	(0.200)
Males	-0.557	-0.938	-0.690	0.347	0.612*	0.069
	(0.607)	(0.487)	(0.355)	(0.252)	(0.294)	(0.225)
Females	0.151	-0.126	0.044	-0.271	0.107	-0.396*
	(0.451)	(0.365)	(0.291)	(0.257)	(0.208)	(0.173)

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 20 and 55 years. Controls used are gender and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2012–2015.

Table A.5: Event study estimates on pre-treatment periods of Figure 9

	Employment			Wages		
	2012	2013	2014	2012	2013	2014
Minimum wage	0.347	0.038	0.094	-0.993	-0.450	-0.734
	(0.356)	(0.255)	(0.204)	(0.926)	(0.863)	(0.536)
40th–50th	-0.607	-0.759	-0.389	0.007	0.477	-0.042
	(0.555)	(0.538)	(0.344)	(0.582)	(0.501)	(0.270)
50th–60th	-0.512	-0.791	-0.593	0.235	0.391	0.105
	(0.580)	(0.433)	(0.319)	(0.292)	(0.391)	(0.198)
60th–70th	-1.019*	-0.813*	-0.496	0.337	0.784*	-0.055
	(0.422)	(0.409)	(0.268)	(0.254)	(0.367)	(0.164)
70th–80th	-0.376	-0.937*	-0.368	0.816**	1.215***	0.252
	(0.443)	(0.365)	(0.228)	(0.259)	(0.320)	(0.284)
80th–90th	0.133	-0.691*	-0.619***	0.464	0.408	-0.009
	(0.370)	(0.299)	(0.178)	(0.737)	(0.534)	(0.181)
90th–100th	0.444	-0.426**	-0.267*	0.103	0.203	-0.345
	(0.720)	(0.161)	(0.121)	(0.333)	(0.358)	(0.380)

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 20 and 55 years. Controls used are interactions of gender with seven age categories and dummy of employee in the base period. Cluster standard errors (G=53). Workers are observed in August of each year. Source: PILA, 2012–2015.

E. Construction of AKM sample

To construct the sample for the AKM estimation, I restrict to six periods 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 observations, after eliminating workers with nonpositive 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 one more than one contribution to the social security system. The nominal wages are transformed to real terms using the monthly CPI from DANE (with base year 2018) and take logarithms of the final expression (lnw_{it}).[A.1] Table A.6 shows descriptive statistics by the seven quantiles of firm fixed effects and Table A.7 shows the decomposition of the variance of wages $Var(lnw_{it})$.

[A.1]The model is estimated using the `reghdfe` command of Stata, the command eliminates 2,507,438 additional singleton observations.

Table A.6: **Descriptive statistics by firm fixed effects**

	Average				
Firm effects by 7 quantiles of $\hat{\psi}_j$	Employment	Sex (%)	Age	Real wages (USD)	N
1	13	0.64	37.7	236	45,883
2	17	0.66	37.3	228	44,212
3	14	0.61	37.5	243	40,764
4	18	0.55	38.0	273	40,675
5	36	0.53	38.9	326	44,176
6	4	0.53	42.6	259	47,912
7	77	0.55	39.5	585	43,233

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 World Bank. Only workers who contribute as employees are taken into account. Source: PILA, 2015-August.

Table A.7: Variance decomposition of lnw_{it}

Share of variance explained by:	
$Var(\alpha_i)$	78%
$Var(\psi_{j(i)})$	16%
$2Cov(\alpha_i, \psi_{j(i)})$	11%
$Corr(\alpha_i, \psi_{j(i)})$	.15

F. Definition of Variables

Formal wages. The nominal contribution to the health system of each worker for August is used. Only positive contributions are taken into account, as zero indicates workers on leave for several reasons that are not related 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.