Essays in Labor Economics: Environmental Shocks, Immigration, and Gender Wage Gap in Linked Employer-Employee Data

by

Hugo Sant'Anna Rodrigues

(Under the Direction of Meghan Skira)

Abstract

This dissertation leverages high resolution data from Brazil to examine how labor markets adapt to major shocks and structural features. It is organized into three chapters that, respectively, investigate: (1) the impact of an environmental disaster, (2) the effects of large-scale refugee inflows, and (3) a new method to decompose the gender wage gap by capturing interactions between workers and firms. Chapter 1 focuses on the 2015 Mariana Dam disaster, offering a comparison between an urban spatial equilibrium model and a factor of production model. I find that the primary labor market disruption stem from the factor of production channel, due to the disaster's profound effect on physical infrastructure. However, traces of spatial equilibrium mechanisms emerge, particularly in certain industries and regions. Chapter 2 examines labor market outcomes tied to the Venezuelan refugee crisis. Employing a doubly-robust difference-in-differences approach, I find that monthly wages for Brazilian workers in Roraima rose by approximately two percent. This effect is concentrated in sectors and occupations with minimal refugee participation. Chapter 3 develops a novel gender wage gap decomposition that accommodates significant "complementarity effects," where the firm premium is conditional on worker characteristics. These complementarity effects account for nearly 17 percent of the observed wage gap, and, combined with the underrepresentation of women in higher-paying, higher-return firms, explain close to half of the total gap.

Index words: Environmental disaster, Immigration, Gender wage gap, Spatial equilibrium, Doubly robustness, Gaussian mixtures, k-Means clustering

Essays in Labor Economics: Environmental Shocks, Immigration, and Gender Wage Gap in Linked Employer-Employee Data

by

Hugo Sant'Anna Rodrigues

B.Sc., Institute of Capital Markets, Brazil, 2017
M.Sc., Baylor University, 2020
M.A., University of Georgia, 2023

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2025

Essays in Labor Economics: Environmental Shocks, Immigration, and Gender Wage Gap in Linked Employer-Employee Data

by

Hugo Sant'Anna Rodrigues

Major Professor:	Meghan Skira
Committee:	Gregorio Caetano
	Brantly Callaway

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2025

Acknowledgments

First, I thank God for making this possible.

My deepest gratitude goes to my family, who invested so much in my education and shaped me into who I am today.

I thank my soulmate, Min Jeong Kang, for inspiring me and providing unwavering support throughout this journey.

I am deeply grateful to Bernardo Aquino, my lifelong friend who also pursued a Ph.D. in the United States, for our meaningful conversations and invaluable guidance.

I thank all the professors and teachers who guided me along this path of education and self-improvement. Special thanks to Dr. Ian Schmutte, who provided life-changing lessons and helped me become an independent researcher. He believed in me even when I did not, and his mentoring and valuable conversations over the past five years as my doctoral advisor have been instrumental.

I am thankful to my professors and committee members, Dr. Gregorio Caetano and Dr. Brantly Callaway, as well as other members of the Econometrics Reading Group, for their insightful suggestions and opportunities. Special thanks to Dr. Scott Cunningham, who has believed in me since I was his master's student and continues to support me to this day. I also thank Meghan Skira, Josh Kinsler, and Roozbeh Hosseini for all their useful comments.

Special thanks also to my friends Samyam Shrestha, Haewon Oh, Miranda Gomez, and Pedro Savarese for their collaboration, creativity, and stimulating philosophical discussions.

Finally, I am grateful to the University of Georgia, particularly the Department of Economics, for providing the administrative and financial support that made my studies possible.

Contents

List of Figures

List of Tables

Chapter 1

Labor Market Effects of an Environmental Disaster: Evidence from the 2015 Mariana Dam Failure

1.1 Introduction

In the recent past, economies across the globe have witnessed severe disruption due to environmental disasters of both natural and anthropogenic origin. These calamities often cause widespread, long-term changes in the affected regions, leaving indelible marks on their landscapes, altering spatial equilibrium, and impacting various economic dimensions, from local labor markets to international trade. Disentangling the multifaceted effects of such incidents is challenging, especially given the vast heterogeneity across such disasters and their consequences.

This study exploits the unique circumstances surrounding the Mariana Dam disaster in Brazil to understand these interconnected effects. A key aspect setting apart the Mariana disaster from other similar incidents is its geographically concentrated impact but extensive spillover. The disaster's epicenter was Bento Rodrigues, a small district in the municipality of Mariana, Minas Gerais, Brazil, with a population of 600 inhabitants. The impending dam rupture prompted a swift relocation of these residents, averting a potential human catastrophe on a larger scale. Therefore, only 19 fatalities occurred despite the total submersion of Bento Rodrigues. Nevertheless, the aftermath of the Mariana

Dam rupture saw the severe contamination of the Doce River, a vital water source and livelihood provider for approximately two million individuals in Minas Gerais and Espírito Santo states. By December 2015, around 55 million cubic meters of toxic iron tailing waste had flowed downstream in a 663.2 km course, contaminating the river ecosystem and its 80 km radius surroundings before reaching the Atlantic Ocean in Espírito Santo. This ecological disaster has been tagged the worst in Brazilian history and the most severe globally involving mining operations. The peculiar nature of the Mariana disaster - immense environmental devastation alongside minimal human capital loss - provides an exceptional lens for examining the economic repercussions of such events.

Central to this investigation is an exploration of the labor market implications of this disaster. Specifically, this research aims to answer how the labor market responds to drastic environmental or climate changes in the face of minimal human capital losses and how individuals adapt to such significant shifts in their environment. Because of the very limited physical destruction of property and human life, I can separate the heterogeneous nature of the disaster and understand how the labor market responds to such changes.

This research is guided by two major schools of thought that appear to provide contrasting views. The first perspective draws on the principles of classical microeconomics, which conceptualizes the environment as a natural resource integrated into the production function as a form of capital. Consequently, negative shocks to such "natural capital" would be expected to drive wages and employment downward due to the quasi-complementarity of human and physical production components.

In contrast, the second perspective emerges from urban economics theory, particularly the Rosen-Roback model of urban spatial equilibrium. This model postulates that natural resources, such as a river, serve as amenities, enriching the quality of life and increasing utility generated by the inhabitants of the region. An exogenous negative shock to such an amenity, such as severe contamination, reduces individual utility levels. Consequently, firms may need to offset this decrease in utility by offering higher wages to retain their workers. Workers may consider relocating to regions offering better living conditions if the compensation is insufficient.

The Mariana disaster presents a unique opportunity to untangle these effects. The unique subtlety of the catastrophe lies in the toxic tailings' dual potential: on the one hand, they could impair the production function of industries relying on the river's resources, and on the other hand, they could pressure firms to increase wages as compensation for the depreciated amenity

value of the region. The balance of these dynamics and their impacts on the labor market forms the core of this study.

This study employs a difference-in-differences model using a rich administrative panel dataset on the universe of formal workers in Brazil to quantify these effects. The approach compares municipalities directly impacted by the disaster with sufficiently similar yet unaffected ones. The primary focus is on the Mariana municipality itself as the treatment sample and the neighboring municipalities unaffected by the disaster serving as controls. The findings indicate a roughly 5.5 percent decrease in aggregate wages, a result that survives across all model variations. Interestingly, this wage reduction trend mirrors the heterogeneous effects observed in the agriculture and mining industries, which generally rely heavily on rivers and water, albeit at a lesser magnitude.

The observed effects extend to other regions significantly impacted by the disaster, namely municipalities through which the Doce River passes and municipalities in Espírito Santo's coastal area, although the impact varies in magnitude.

The study also investigates job retention and the probability of relocation by incorporating these aspects into a linear probability outcome in the DID model. The findings here offer limited insights, yet the overarching implications suggest that drastic environmental or climatic changes that do not directly threaten human life can still detrimentally impact the market's production function and reduce the population's overall wealth.

Nonetheless, the study also points out the presence of Rosen-Roback effects, that is, compensatory wage increases, particularly among worker groups whose roles are not directly dependent on water as a production component, such as those in office occupations. While these effects are indeed present among the potential market responses in the wake of the disaster, they appear neither strong nor representative enough to influence the aggregate level significantly.

My study contributes to the literature on the effects of disasters on labor market outcomes and urban spatial equilibrium. A good example of a disaster being studied is Hurricane Katrina, that devasted New Orleans in 2005. Groen and Polivka (2008) compares evacuees and individuals unaffected by the hurricane to find that initially, those affected suffered adverse labor market conditions but recovered quickly in the following months, especially if the individual decided to return to the region where the disaster occurred. Other studies also point out room for a limited recovery of some sectors in the affected regions and a "bounce back" behavior of labor market experiences of individuals (Vigdor, 2008; Zissimopoulos and Karoly, 2010). McIntosh (2008) finds that the

migration from New Orleans to Houston negatively affected the labor market outcomes of Houstonians.

Other studies have explored the effects of droughts on the labor market, finding severe adverse effects on women's workdays and mobility (Efobi, 2022; Afridi, Mahajan, and Sangwan, 2022).

On the other way, Kirchberger (2017) explores an earthquake that devasted the coastal region of Indonesia to find a resilient job market with positive wage growth driven by a limitation in labor supply in rural areas.

From the urban economics perspective, studies such as Frame (1998), Ortega and Taşpınar (2018), and Boustan, et al. (2020) explore variations in the environment, such as rising sea levels and flooding, to suggest the adverse effects of disasters on spatial equilibrium outcomes.

I fill the gap in the literature by exploring the interplay between capital shocks and spatial equilibrium dynamics, leveraging the unique nature of the Mariana Disaster. For instance, as a natural occurrence, the Katrina hurricane did not transform the environment to adverse potential toxic conditions. Moreover, the scale of the hurricane effectively destroyed a major city in the U.S., irrevocably mixing severe human capital shocks with changes in amenities. A recent accident that draws an interesting parallel to the Mariana Disaster is the East Palestine train derailment in Ohio, which in February 2023 released an enormous amount of toxic fumes into the atmosphere, particularly vinyl chloride, a substance supposedly carcinogenic to humans (Schnoke, et al., 2023). Moreover, to the best of my knowledge, this is the first study of the Mariana Disaster that employs rigorous econometric models to capture its causal effects on the labor market using heavily detailed data.

The remainder of the paper is organized as follows: Section 3.2 provides the background and reports on the government and other agents' reactions to the incident. Section 3.3 describes and explores the data used in my study. Section 1.4 is reserved for the identification strategy of the research and spatial data exploration. Section 3.4 provides the empirical framework. Section 3.5 and Section 3.6 present the main result together with its robustness checks and discusses the underlying mechanisms. Finally, Sector 3.7 concludes. I reserve Appendix Section A.1 to present the theoretical framework of Rosen-Roback spatial equilibrium and Factor of Productivity models.

1.2 Background

The Mariana disaster, which took place on November 5, 2015, is widely recognized as one of the most devastating environmental catastrophes in Brazilian

and world history. Triggered by the collapse of the Fundão tailings dam, owned by Samarco (a joint venture between Vale S.A. and BHP Billiton), the disaster released an estimated 45 to 60 million cubic meters of iron ore extraction waste. Among other residues present, high mercury and other toxic heavy metal concentrations were found, generally used for mining operations (Hatje, et al., 2017). The ensuing massive mudflow obliterated the small district of Bento Rodrigues, part of Mariana municipality in Minas Gerais.

The disaster's impacts extended far beyond the immediate vicinity of the dam. The mudflow traveled approximately 650 kilometers along the Doce River, reaching the Atlantic Ocean 17 days after the dam collapse (Hatje, et al., 2017). This had significant implications for the coastal environment and marine life (Gabriel, et al., 2020). In particular, the Espírito Santo region, known for its coastal fishery activities, was severely affected. The influx of iron ore waste led to a dramatic increase in water turbidity and heavy metal presence in marine life, which disrupted the photosynthesis process for aquatic plants and corals, ultimately reaching fish populations.

The Brazilian government's response to the disaster was multifaceted. Instants right before the disaster, it was issued an evacuation order for Bento Rodrigues' 600 inhabitants, effectively saving all but 19 lives. As the mud flowed, water distribution had to be interrupted, with the municipalities diverting water from other areas to supply their inhabitants. The government also issued a ban on fishing near the Doce River estuary and along the affected Atlantic Coast, potentially disrupting several markets.

Initially, Samarco, the company responsible for the incident, was fined approximately 66 million 2015 dollars. In March 2016, a settlement was reached in which Samarco, Vale, and BHP agreed to pay 5.3 billion dollars over 15 years to restore the environment and communities affected by the disaster completely. Recovery is estimated to take several decades (Fernandes, et al., 2016).

Despite its catastrophic scale, the disaster resulted in a surprisingly minimal loss of human life. The municipality of Mariana, which served as the epicenter of the disaster, is home to approximately 50,000 inhabitants. Remarkably, the majority of this population was spared, preserving a significant portion of the region's human capital. However, the environmental devastation inflicted upon the area was profoundly severe and, arguably, irrecoverable.

The disaster in Mariana provides a unique opportunity to investigate the interaction between a shift in the spatial equilibrium and labor market dynamics within the same market. Specifically, how external environmental shocks influence employment, wages, and worker mobility patterns. This insight is

crucial for informing policymakers and preparing for future environmental disruptions.

1.3 Data

The data used for this study is the Annual Registry of Social Information (RAIS), an administrative panel data maintained by the Brazilian Ministry of Labor, containing information on the universe of the Brazilian formal labor market. I focus on the period from 2008 to 2018. Because the disaster occurred right at the end of the year of 2015, I consider, for the sake of simplicity, the year 2016 to be the first treatment year of my analysis.

The dataset provides a detailed description of its worker base. For instance, it identifies the individual through the Brazilian equivalent of a social security number, called PIS (Social Integration Program in Portuguese). Several labor market outcome determinants are present: gender, race, age, tenure in months, individual's education level, and nationality. Moreover, I observe the worker's occupation according to the Brazilian Occupation Code.

Other variables of interest are the type of work, which tells the employer-employee contract nature (if it is temporary or not). Workers in Brazil can have part-time or full-time jobs, expressed in the variable working hours per week. There is also the worker's hiring date, separation date, separation cause, and if the worker was present on December 31st of the observed year, the main variable that tells if a worker lost their job or not.

One of the key variables in my research is the employee's municipality of work, a strong indicator of their residency. The RAIS dataset also provides the worker's firm's identification code, CNPJ (National Registry of Legal Person, in Portuguese). CNPJ is strongly correlated with the municipality. A company with two branches in different locations appears as two separate CNPJs in the dataset.

Another crucial variable in RAIS is the economic activity code, CNAE (National Registry of Economic Activities, in Portuguese). CNAE code is a highly detailed categorization of the firm's main economic activity. Therefore, I can discriminate firms and groups of workers based on their ultimate production function's output. The code comprises seven numbers, each increasing a degree of detail. For instance, "01" represents agriculture activities, while "0155505" represents agriculture activities related to poultry, egg production.

1.3.1 Generating Outcome Variables in RAIS

The richness of RAIS allows for a comprehensive exploration of labor market dynamics in the aftermath of the disaster. To guide this investigation, I focus on three primary labor market outcome variables: average monthly wage, a "mover" indicator, and a dismissal indicator.

The core variable in this study, "average monthly wage", measures the typical monthly earnings a worker garners in a year. As is common with wage data, outliers may skew the analysis, and imputation errors may be present. To counteract these concerns, I narrow the scope to include only full-time workers working more than or equal to 20 hours weekly. Further, I apply a winsorization technique at the 2.5 and 97.5 percentiles to mitigate the impact of potential wage distortions. This method replaces the extremes of wage distribution with the nearest values within the defined percentiles, thereby reducing the influence of wage extremities or imputation errors on the final results.

Given the prevalence of multi-job holding among Brazilians, particularly those earning low or minimal wages, I ensure that the analysis pertains solely to primary jobs. I define the primary job as the one with the earliest hire date for an individual, as indicated in the dataset. In rare instances where multiple jobs share the same hire date, I default to the job with the highest pay. This data preparation procedure ultimately creates roughly individual-year-municipality spells.

The "dismissal" variable is a binary indicator tracking whether a worker maintained employment with the same firm until December 31st of a particular year. The variable assumes a value of one if the worker separates from the firm before the year-end; otherwise, it is zero.

The "mover" variable, developed by longitudinally tracking workers using their PIS, identifies individuals who change their work municipality across two consecutive years. Constructing this variable entails three steps. First, I identify individuals present in the research's region of interest in 2008. Then, I trace these individuals across all subsequent years, irrespective of their location in Brazil. This process repeats annually until 2018.

The second step involves examining individuals annually and noting any changes in their municipality of work. For each PIS, I compare the current municipality with the municipality in the following year, sorted by year and hire date. A change in municipality code leads to the assignment of a value of one to this new variable, while a lack of change results in zero.

The final step involves pruning observations not situated within the region of interest. As such, the effects quantified in this research should be understood as "local market effects." This approach prioritizes the original location over

individual identifiers, thereby ensuring that the analysis faithfully reflects the local labor market's dynamism.

My sample must also provide a basis for comparison across control and treatment groups. This implies ensuring a balance in all labor market outcome determinants employed in my study. In other words, I only consider instances where I observe at least one corresponding element in both cohorts. Such an approach is crucial for mitigating bias and ensuring the reliability of my analysis. I further elaborate on the nature of such determinants and how I deal with panel data imbalance in the sections dedicated to the identification and empirical strategy.

1.4 Identification Strategy

To estimate the causal effect of the Mariana disaster on labor market outcomes, I employ a difference-in-differences (DID) approach, which levers the variation in treatment status across time and between affected and non-affected municipalities.

The main area of interest is the region called *Quadrilátero Ferrífero* (roughly translated as Iron Square). The Mariana municipality, where the disaster took place, belongs to this region characterized by mining operations and its historical importance.

I exploit the fact that Doce River starts from Mariana and outflows the region, as shown in Figure 1.1. The brighter area corresponds to the Mariana Municipality, while the darker areas are the other municipalities belonging to the Quadrilátero Ferrífero region, which I use as the control group. The dark thicker line going outward from Mariana is the affected river. Given the event happened in November 2015, I consider 2015 the reference and 2016 until 2018 the post-treatment period. I also employ data from 2008 until 2014 as additional pre-treatment periods. It is possible to measure the causal effects using this method because the municipality where the individual is located is observed, together with the individual identifier. However, there are some concerns that arise with this identification strategy.

Although not as urban as the state capital metropolitan region, the region itself is sufficiently integrated to generate spillover effects due to individuals moving across municipalities. After the disaster, affected workers may have decided to move to locations and jobs that I use as the control, potentially distorting any measurement when employing the two groups. I address this issue with fixed effects interactions between social identifiers and the location of the individual. When spells related to the individual-municipality relationship are

included in the model, the parameters absorb any effect related to geographical movement. Additional details are provided in Section 3.4.

Moreover, two effects may be at play when measuring the Mariana Disaster. The disaster itself, in other words, the dam rupture and the water contamination that eventually took over the Doce River. Separating these effects in this context is challenging, particularly when I ultimately intend to disentangle TFP and Rosen-Roback effects. I address this issue by employing the empirical analysis in two other regions affected by the disaster: the continuation of the Doce River until the Minas Gerais border and the river estuary region in the Espírito Santo state. Further explanation is provided in Section 3.6.

Lastly, the dataset employed is highly unbalanced by nature, given that individuals may leave the labor market at any moment during the studied period, creating potential overlapping issues between treatment and control individuals. To counter this issue, I employ a doubly-robust regression where propensity score weights are used to balance both groups. I discuss this method in Section 3.4.

For a summary statistics of the region, refer to Appendix Section A.2

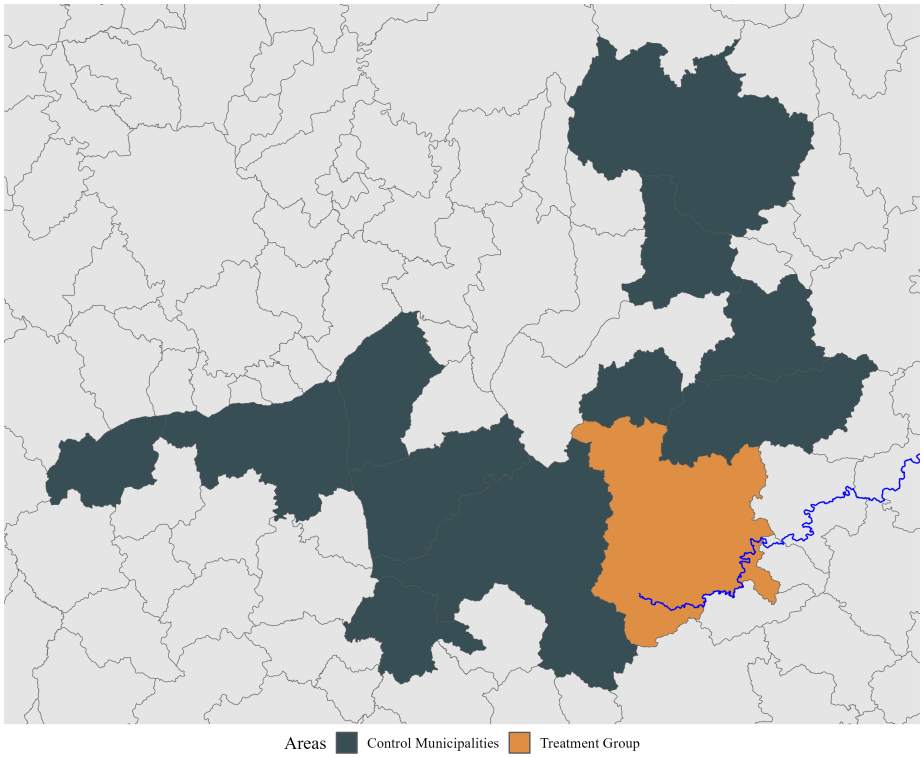


Figure 1.1: Overview of the Iron Square Region and the Doce River

1.5 Empirical Strategy

The empirical strategy used in this paper exploits the quasi-random nature of the disaster to design a natural experiment setting. My identification strategy relies on the assumption that conditional on the control variables, the disaster is as good as randomly assigned across municipalities and individuals within the municipalities. I use a difference-in-differences approach to estimate the causal effect of the disaster, comparing the changes in labor market outcomes in the affected municipalities before and after the disaster to the changes in a control group of municipalities over the same period.

My main setting compares Mariana, the municipality where the disaster occurred, with municipalities from the same region of Minas Gerais, called the Iron Square Region[1]. Later, I further explore the disaster by analyzing additional regions in the Discussion Section.

[1] This is a rough translation from *Quadrilátero Ferrífero* in Portuguese

1.5.1 Fixed Effects Models

Here I present the specification for the fixed effects models. The basic version uses a two-way fixed effect regression using social identification and year fixed effects. I also present an alternative design of the baseline model to address potential spillover effects.

Baseline Model

The baseline specification is a fixed effects model that utilizes the individual social identification and year dummies. This model is outlined as follows:

$$y_{imt} = \beta D_{imt} + \mu_i + \lambda_t + \epsilon_{imt} \tag{1.1}$$

where y_{imt} is the labor market outcome of individual i at time t and municipality m, D_{imt} is a treatment indicator that equals 1 if individual i is located in a disaster-affected municipality m at time t and 0 otherwise. The parameters μ_i and λ_t capture time-invariant individual effects and yearly fixed effects, respectively. Lastly, ϵ_{imt} represents the error term that accounts for unobserved model characteristics. The key parameter of interest here is β, which quantifies the causal effect of the disaster.

The study covers the years from 2008 to 2018. As the disaster occurred in November 2015, and it took approximately one month for the waste to reach the ocean, 2016 is considered the first year of treatment for any region specification. The labor market outcomes for individual analysis are logarithmic wages and linear probabilities of being dismissed and moving from a municipality.

Alternative Baseline Model: Interaction Model

There are some potential challenges with the baseline model. Given the geographical proximity of the sampled municipalities and the intricate nature of the local economy, it's possible for individuals to relocate from the control to the treated region and vice versa. Such movement could introduce bias into the baseline model results. I propose an interaction model allowing individual interactions with their municipality to address this issue. This approach should capture the effects driven by changes in location, thereby allowing us to isolate the impact of the disaster on the local labor market. It is based on previous studies such as Sant'Anna and Shrestha (2023) and Foged and Peri (2016). The new model is specified as follows:

$$y_{imt} = \beta D_{imt} + \mu_i + \mu_m + \phi_{im} + \lambda_t + \epsilon_{imt} \tag{1.2}$$

where the municipality fixed effects is μ_m, similarly to the secondary model. The interaction term is ϕ_{im}, where I create individual-municipality spells based on individual i social identifier (pis) and the municipality geographical code from the Brazilian Institute of Geography and Statistics (IBGE).

1.5.2 Augmented Inverse Propensity Score Weighting (Doubly Robust) Models

I also implemented an augmented propensity score weighting procedure to increase the robustness of the fixed effects models. This procedure, also known as the Doubly-Robust approach, combines ordinary fixed effects regression with inverse propensity score weighting. This method addresses potential selection bias stemming from differences in the groups' support of covariates. The advantage of this method is the need for only one procedure to be correctly specified, the fixed-effects regression or the propensity score weighting, to prevent misspecification (Chernozhukov, et al., 2017; Robins, Rotnitzky, and Zhao, 1994).

My approach is similar to Strittmatter and Wunsch (2021), adapted to a difference-in-differences design. The first step involves estimating the propensity score of each individual to belong to the treatment group. This is achieved using a logistic regression model with individual characteristics as the independent variables:

$$p(X_i) = Pr(D_i = 1|X_i) = F(\theta' X_i) = \frac{e^{\theta' X_i}}{1 + e^{\theta' X_i}} \tag{1.3}$$

where X_i denotes a vector of time-invariant covariates for individual i and $D_i = 1$ signifies that the individual belongs to the treatment group, namely

the disaster-affected municipalities. The parameter θ establishes the logistic relationship between the covariates and which group the individual belongs to. In the next step, I predict $p(\hat{X}_i)$, the estimated propensity score, for all observations within the sample. The controls used are age, gender, race, education level, tenure, job occupation, and the firm's economic activity. Weights are then constructed by the following equation:

$$\hat{W}_{it} = \frac{(1 - D_{it})\hat{p}(X_i)}{1 - \hat{p}(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{(1 - D_{it})\hat{p}(X_i)}{1 - \hat{p}(X_i)} \tag{1.4}$$

Ideally, the weights assigned to individuals would remain unchanged across different years. However, given the potential for changes in the treatment and control sample over time, due to movement across municipalities and labor market leavers, I compute a set of weights for each year in the sample.

The final step minimizes the weighted sum of squares. Let $M = \{1, 2\}$ represent the two models specified in Subsection 1.5.1, i.e., the baseline and the municipality-individual spell model, respectively. Let $\boldsymbol{\alpha_M}$ represent the corresponding fixed-effect set used for each specification. Therefore, the minimization problem is framed as follows:

$$\hat{\beta}_{dr}^{M} = \arg\min_{\beta} \sum_{N} w_{it}(y_{imt} - \beta D_{imt} - f(\boldsymbol{\alpha_M}))^2 \tag{1.5}$$

where $w_{it} = \hat{W}_{it}$ is calculated using Equation 4. N is the number of observations in the sample, T is the number of time periods. The function $f(\boldsymbol{\alpha_M})$ represents a linear function wrapping the fixed-effects set:

$$f(\boldsymbol{\alpha_1}) = \alpha_i + \alpha_t \tag{1.6}$$

$$f(\boldsymbol{\alpha_2}) = \alpha_i + \alpha_m + \alpha_{im} + \alpha_t \tag{1.7}$$

As customary with difference-in-differences approaches (MacKinnon, Nielsen, and Webb, 2023), I cluster my standard errors at the municipality level, which is the dimension in the data that best represents the disaster's geographical level of impact.

1.6 Main Results

In this section, I present the main results using the Mariana municipality as the treatment group, comparing similar municipalities in the "Iron Square" region.

Table 1.1: Main Results for Mariana Region Analysis

	Fixed Effects Models		Doubly Robust	
	(1)	(2)	(3)	(4)
Log Wage: Treat x Post	−0.066***	−0.052***	−0.054***	−0.055***
	(0.016)	(0.010)	(0.013)	(0.009)
Dismissed: Treat x Post	−0.003	−0.005	0.041	0.010
	(0.036)	(0.021)	(0.026)	(0.021)
Individual FE	X	X	X	X
Municipality FE		X		X
Ind. x Municipality FE		X		X
Year FE	X	X	X	X
N Clusters	12	12	12	12
N	1 567 721	1 567 721	1 567 721	1 567 721

[1] Standard-errors are clustered by municipality.
[2] * p < 0.1, ** p < 0.05, *** p < 0.01
[3] Covariates used for propensity score estimation in DR models are age, tenure, education level, gender, race, worker's occupation, and the firm's main economic activity.

Table 2.3 provides the regression results for the baseline and alternative baseline fixed effects models in Columns (1) and (2), respectively, and their Doubly Robust versions in Columns (3) and (4).

In terms of the Fixed Effects model, all specifications provide significant results at the 1% level or lower for the effect of the disaster on wages (Log Wage: Treat x Post). The coefficient estimate values at -0.066 for the baseline in Column (1) and -0.052 when adding the interactive individual-municipality fixed effects parameter in Column (2), suggesting a notable decrease in wages after the disaster in the treated areas.

The results' significance and direction are insensitive to the DR regression set, with -0.054 in Column (3), representing the doubly robust version of the baseline model, and -0.055 for the DR version of the interacted fixed effects model, all at the 1% significance level.

Even though the log-wage study provided significant evidence of negative effects, the effect on dismissals (Dismissed: Treat x Post), however, appears insignificant in all model specifications, with most magnitudes being negligible except for the doubly robust baseline in Column (3), indicating that the disaster did not significantly alter the dismissal rate in the treated municipality of Mariana, at least when comparing to the neighboring, non-affected, municipalities.

These results provide strong initial evidence of a substantial negative impact on wages due to the Mariana Dam disaster in its own municipality. One could be inclined to affirm that, given the negative nature of wage effects, it was, in

the first place, a capital shock in accordance with the Factor of Production hypothesis. Nevertheless, these results are not sufficient to disentangle the Factor of Production and Rosen-Roback effects. There is concern about the disaster taking place in the region, creating massive destruction of capital when the dam was destroyed, and potentially devastating mining operations or economic activities surrounding the incident area. Therefore, these results capture not only the total contamination of the river but also the disaster itself.

1.6.1 Robustness Checks

Before proceeding with the mechanisms and heterogeneous effects underlying the results, I discuss the survivability of my procedure when applying typical robustness checks used in difference-in-differences designs. In this section, I investigate whether the results yielded by the previous log-wage regressions are due to randomness in the control municipalities.

Event Studies

The event study methodology serves to validate the common trends assumption, a fundamental prerequisite for employing the Difference-in-Differences (DID) identification strategy (Callaway and Sant'Anna, 2021b; Goodman-Bacon, 2021). In essence, this assumption states that, in the hypothetical scenario where no treatment was administered, both the control and treated groups would exhibit parallel trends over time. However, the challenge arises when examining post-treatment periods, as the counterfactual outcome is unobservable. Nevertheless, I test the pre-treatment trends to make sure the null hypothesis, that the control and treated groups exhibit the same estimate trend, is not rejected.

Transforming the event study method into a regression model allows us to dissect the treatment effect parameter across distinct time periods. This strategy requires the identification of a reference period, ideally just before treatment, to compare the dynamic estimates. In this study, I choose 2015 as the pre-disaster reference year, primarily due to the timing of the event. The disaster unfolded in November, taking less than a month to reach the ocean, which suggests that the bulk of 2015 remained unaffected by the dam rupture. Consequently, 2015 is regarded as the reference year.

If the impact took some time to manifest in the labor market, we would expect it to become apparent in the subsequent year, affirming my specification's validity. On the other hand, if the shock was immediate, my choice could be seen as a 'conservative' approach. This method would likely result in estimates bi-

ased towards zero, as we are potentially incorporating part of the post-treatment period into our pre-treatment reference period.

The empirical specification is as follows:

$$y_{imt} = \sum_{t=2008}^{2018} \beta_t D_{imt} + f(\boldsymbol{\alpha_M}) \quad \text{for} \quad t \neq 2015 \tag{1.8}$$

where β_t is the decomposed estimate parameter of interest. $f(\alpha_M)$ represents a linear function with the corresponding fixed-effects model from equations 1.6 and 1.7. I also present event study versions for the doubly robust alongside the two baseline models, further reinforcing my robustness checks.

Figure 1.2 presents the plot panel of all models. The first column represents the baseline version, while the second column shows regressions from the doubly robust specification. The first row is the simplest model where only individual fixed effects are included to control for time-invariant characteristics. The second row represents the model with the individual and municipality interaction terms.

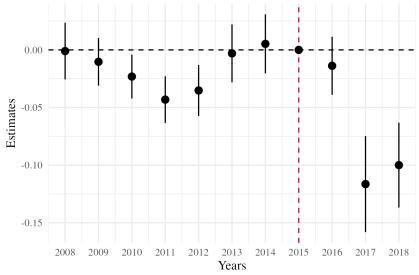
In general, all results are robust with different degrees of success. The trends do not change dramatically with different settings. Notably, in any model, there is no difference between the reference year and the estimate in 2016. There are possible explanations behind this phenomenon. The immediate effect could have taken some time to appear in the aggregate market. There could also be conflicting heterogeneous effects between two industries that would absorb the shock differently, such as healthcare and the mining industry. Nevertheless, the result that appears in the following years indicates the lasting effects in the aggregate market are related to capital shocks, in accordance with the Factor of Production hypothesis.

Placebo Tests

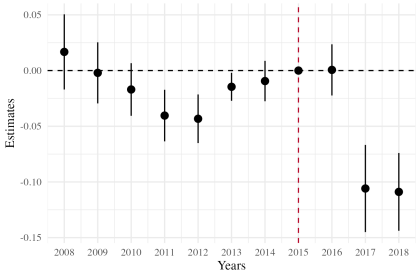
The effects observed in the regressions could be driven by random variations in the control group. Given Mariana is the sole treated region in the main setting, with another eleven municipalities serving as control, it is a natural assumption that perhaps the observed estimates are generated by some municipalities having disturbances after the disaster.

To address this issue, I elaborate a permutation test where I eliminate Mariana from the sample and permute each municipality as if they were treated instead. This is similar to the robustness approaches proposed in synthetic control methods, such as Abadie, Diamond, and Hainmueller (2015), Abadie,

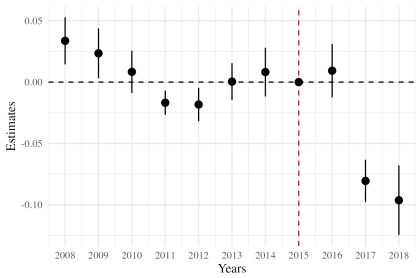
Figure 1.2: Event Studies for Log Wage Regression Models



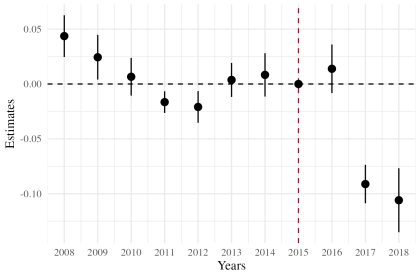
(a) Individual Fixed Effect ES



(b) Individual Fixed Effect ES - DR



(c) Ind. x Municipality Fixed Effect ES



(d) Ind. x Municipality Fixed Effect ES - DR

Diamond, and Hainmueller (2010), and Abadie and Gardeazabal (2003). Figure 1.3 shows the placebo test result.

The horizontal axis represents the municipality code used as the placebo treatment for my most "robust" model, corresponding to the DR design with municipality-individual spells in the fixed effects parameters. The dashed vertical line represents Mariana's estimate, the original treated group. When performing the placebo tests, I remove Mariana from the sample and test the selected placebo with the remaining municipalities. Ideally, all other municipalities' results would be centered at zero, with Mariana being by far the most affected and isolated. However, the result reveals that seven out of eleven municipalities reject the null hypothesis. Even though it has some degree of undesirability, looking carefully at the estimates, Mariana still has the largest magnitude, in accordance with the fact it is where the disaster occurred. Moreover, the placebos are, in the majority, negatively biased, meaning that my original main result is potentially biased toward zero.

The most problematic municipalities are 310230 and 314480, Alvinópolis, and Nova Lima, with positive estimates that could negatively drive the main

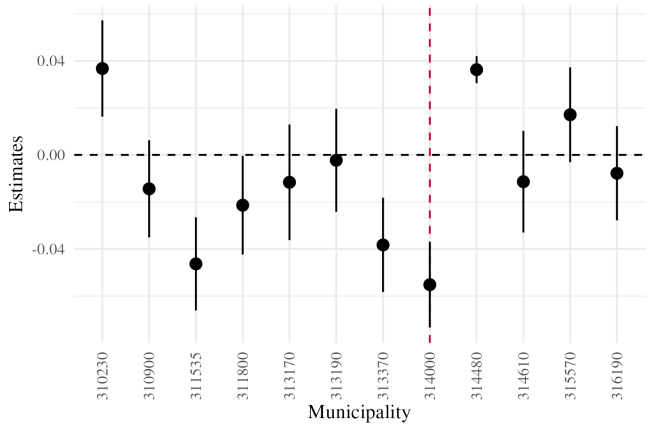


Figure 1.3: Permutation Tests for Ind. x Municipality FE in a Doubly Robust Design

results even further. However, the two municipalities do not have sufficient representatives in the sample to drastically change the effect.

The placebo analysis allows the conclusion that a considerable proportion of the effects measured by the main regressions are driven by the Mariana disaster, not some randomness happening in the control municipalities.

1.7 Discussion

This section discusses the mechanisms underpinning the principal findings of this study. Initially, I explore the Factor of Productivity hypothesis through a heterogenous effect approach, specifically focusing on industries strongly connected to water usage and dam operations. Moreover, I also test the healthcare labor market, as it serves as an illustrative case to ascertain a positive labor demand alteration after the accident that is unrelated to Rosen-Roback effects. There is some evidence that individuals related to clinics and hospitals registered a positive change in wages, presumably due to the surge in labor demand triggered by health-related contingencies in the local populace.

The subsequent segment investigates the influence on the extended region. Using spatial data obtained from the Institute of Brazilian Geography and Statistics, the state of Minas Gerais is partitioned into its respective water basins. This allows for a comparative analysis between municipalities in the Doce River water basin directly impacted by the catastrophe and those in unrelated water basins, with particular care taken to exclude Mariana or the municipalities in the Iron Square from the sample. The main objective of this secondary study is to dis-

entangle the effects of the dam rupture itself from the consequences of water pollution. This distinction serves to highlight the dynamics between Factor Productivity and Rosen-Roback models in the context of severe environmental perturbations.

In the final portion of this section, I focus on "pure Rosen-Roback effects" observed in certain occupations and economic activities. While these effects are discernible at the micro level, there is no evidence of it at the aggregate market level, aligning with the insights thus far learned from previous results, where individuals, on the aggregate market, absorb negative shocks in wages when facing a dismal change in environment.

1.7.1 Verifying the Factor of Production Hypothesis

The Factor of Production Hypothesis states that water is a component of the production function of Mariana and, therefore, when the dam rupture contaminated the river, it disrupted several economic activities in the region that use water in their framework. The main industries with potential affected outcomes are mining, agriculture, forestry, and fishing.

To measure the heterogeneous effect of these industries, I interact the treatment indicator function with another dummy variable that values one when the individual works in the aforementioned industry. I show the modified baseline models in the following equation:

$$y_{imt} = \beta D_{imt} \times I_{if} + f(\alpha_M) + \epsilon_{imt} \tag{1.9}$$

where I_{if} is the new indicator function for when firm f of worker i belongs to the industry of interest. I use the economic activity code variable in RAIS to identify iron mining, agriculture, forestry, and fishing. The latter three activities, however, are not sufficient for testing on their own, and therefore I aggregate them into "rural activities". Iron mining operations correspond to the 07 code, while agriculture, forestry, and fishing correspond to 01, 02, and 03, respectively.

I also provide an additional test where I measure the impact of the disaster on healthcare workers. This is another robustness check to see if the results follow the main intuition. Even though the disaster was minimal in direct human capital loss, the environmental destruction and water contamination are catalysts for negative mental health and well-being shocks. Therefore, healthcare workers are expected to experience an increase in wages or negative effects on dismissals, driven by the sharp increase in the market's labor demand. The corresponding code for healthcare firms is 86, 87, and 88.

Water as capital: Rural and Mining Activities

Table 1.2 presents the results of the heterogeneous effects by economic activities, particularly focusing on iron mining, rural activities (agriculture, forestry, and fishing), and the healthcare industry, which are likely the most affected by the event. I provide the fixed-effect baseline models in the first two columns, while in the last batch, I incorporate propensity score weights, as before, for the doubly robust framework. Each column represents a different fixed-effect approach in the same fashion as the main results model. Each row represents a heterogeneous effect based on the specified industries.

Table 1.2: Heterogeneous Effect Analysis by Economic Activities for Mariana Region

	Log Wage			
	Fixed Effects Models		Doubly Robust Models	
	(1)	(2)	(3)	(4)
Iron Mining: Treat x Post	−0.196***	−0.297***	−0.240***	−0.294***
	(0.017)	(0.009)	(0.028)	(0.005)
Rural Activities: Treat x Post	−0.063**	−0.055***	−0.025	−0.025
	(0.023)	(0.011)	(0.023)	(0.020)
Healthcare: Treat x Post	0.080***	0.074***	0.090***	0.085***
	(0.015)	(0.012)	(0.019)	(0.023)
	Linear Probability of Moving			
Iron Mining: Treat x Post	0.081	0.075**	0.141***	0.122***
	(0.052)	(0.024)	(0.017)	(0.021)
Rural Activities: Treat x Post	−0.011	−0.028	0.018	−0.001
	(0.031)	(0.022)	(0.014)	(0.016)
Healthcare: Treat x Post	−0.082*	−0.086***	−0.046***	−0.064***
	(0.041)	(0.022)	(0.013)	(0.015)
Individual FE	X	X	X	X
Individual x Municipality FE		X		X
Municipality FE		X		X
Year FE	X	X	X	X
N Clusters	12	12	12	12
N	1 567 721	1 567 721	1 567 721	1 567 721
N (Mover)	1 348 847	1 348 847	1 348 847	1 348 847

[1] Standard-errors are clustered by municipality.
[2] * p < 0.1, ** p < 0.05, *** p < 0.01
[3] Covariates used for propensity score estimation in DR models are age, tenure, education level, gender, race, worker's occupation, and the firm's main economic activity.
[4] Iron Mining activity code is 07. Rural Activities correspond to agriculture (01), forestry (02), and fishing (03). Healthcare code is 86, 87, and 88.

Iron mining, being the primary industry directly linked to the disaster, shows a strong, negative wage effect across all models. Notably, the baseline models in Columns (1) and (2) reveal a negative, statistically significant effect of the treatment after the disaster, ranging from -0.196 to -0.297. These figures indicate a substantial reduction in wages in the post-disaster period, potentially reflecting the direct impacts of the disaster on this industry, including the suspension of operations, workforce reduction, lowered productivity, and the loss of the dam itself.

This observation is mirrored in the doubly robust models in Columns (3) and (4). In this case, the estimated wage effects are slightly more pronounced, ranging from -0.240 (Column 3) to -0.294 (Column 4). In other words, there was at least a 20 percent decrease in wages in the mining industry in the disaster aftermath, with a considerable amount of this magnitude directly related to it.

In contrast, rural activities exhibit another, albeit similar, pattern. While the wage effect remains negative, the magnitude is markedly smaller compared to that of the iron mining industry. The fixed effects models suggest an effect ranging from -0.063 (Column 1) to -0.055 (Column 2), while the doubly robust models give an estimate of -0.025 (Columns 4 and 6). However, the DR specification does not yield statistical significance. Given that the accident happened in a small specific region (the district of Bento Rodrigues), and I measured the entire municipality, not only the mentioned district, these results are the first evidence of a negative capital shock related to the pollution, not the dam destruction. Still, the significance level reveals that the results yielded may be driven mainly by outcome determinants imbalance between control and treatment groups.

The disaster as a catalyst for labor demand: The healthcare service

Table 1.2 also provides results related to healthcare sector wage variations, showing a significant increase in wages post-treatment periods across both the fixed effects and doubly robust models. This is likely due to the surge in demand for healthcare services following the disaster, requiring a bolstering of the healthcare workforce in the region and consequently leading to wage increases.

Nevertheless, the possibility exists that these positive outcomes are reflections of "Rosen-Roback" effects, propelled by compensatory wage adjustments in response to shocks to local amenities. This hypothesis can be examined more closely by assessing the probability of migration. By tracking the same individual across time periods or even within the same time period but in distinct jobs, we can determine if their subsequent location differs from their current one. If a change in location is identified, a value of one is assigned to this dummy

variable; if not, it is assigned a zero. This approach provides insight into the mobility patterns of individuals in the wake of such environmental disasters and helps discern if the wage increases are indeed an outcome of Rosen-Roback effects.

The regression models are the same as those used for log-wage outcomes. The outcome of interest, in this case, is the dummy variable representing whether the individual moved in a linear probability model fashion. Results are presented in the second set of regressions in Table 1.2. Results for the mining industry suggest the probability of an individual leaving Mariana after the disaster increased from 7.5 to 14.1 percent, in accordance with the economic intuition of individuals leaving due to severe disruption in production. The precision of the results increases when employing the doubly robust procedure. For agriculture, estimates did not yield any significant measurement across all models.

For the healthcare industry, all results are negative, with varying degrees of significance, revealing that the probability of moving for a healthcare worker decreased after the disaster. The negative direction of these estimates shows the ambiguity of the effects. It could be influenced by a surge in labor demand due to the health shock caused by the disaster or as a consequence of the compensatory wages at play.

1.7.2 Extended Regions Analysis

The Mariana region sample inherently intertwines the immediate disaster and its ensuing environmental pollution. As such, the results derived so far can be regarded as a composite of capital destruction and subsequent environmental fallout at best. To scrutinize human behavior in the face of disruptive environmental changes with minimal direct human capital loss, I turn my attention to regions solely affected by the pollution of the Doce River and the Atlantic Ocean. If these regions exhibit findings akin to those of the Mariana region, it will serve as further evidence supporting the proposition that the overall behavior of the Factor of Productivity hypothesis tends to eclipse Rosen-Roback effects in such circumstances. Figure shows the new regions explored.

The empirical strategy applied is the same as before, fixed effect models augmented with propensity score weights. However, due to the different geographical locations, I propose an alternative version of the identification strategy for the two additional samples, albeit still maintaining the difference-in-differences approach.

Identification Strategy for the extended region within Minas Gerais

The first extended region under study encompasses municipalities situated along the Doce River but external to the Quadrilátero Ferrífero region. This is visually depicted in Figure 1.4a, wherein the brighter highlighted municipalities represent the treatment group adjacent to the polluted river.

I exploit the fact that the Doce River serves as the primary water body of its water basin to select the control group. Using spatial data from the Minas Gerais government for the municipality and water basin borders, I am able to identify municipalities with their water system unrelated to the disaster area. Specifically, I choose the Paranaiba River and Grande River water basins. They are situated diametrically opposite the Doce River and are not directly connected, ensuring that there is a sufficient sample of individuals comparable to the treatment group and a similar climatic backdrop conducive to analogous economic activities in both groups, but also making sure these municipalities will not suffer from spillover effects due to the Mariana Disaster.

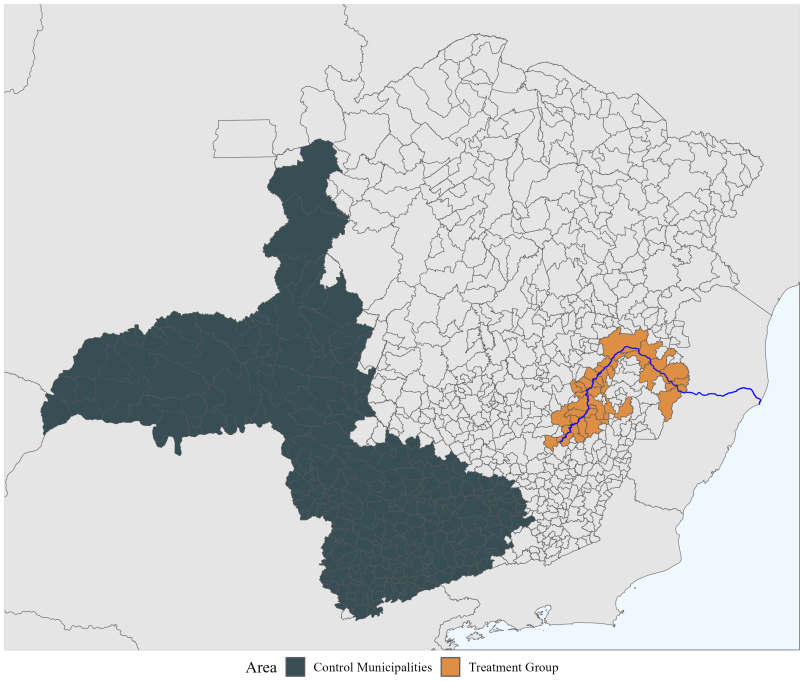
Identification Strategy for the Espírito Santo State

The disaster's final region of impact was Espírito Santo state, the convergence point of the Doce River and the Atlantic. By December 2015, pollution had navigated the river's course, culminating in the contamination of the ocean. This event resulted in a government-imposed prohibition on local fishing activities, a considerable repercussion given the substantial number of coastal fishermen in the area.

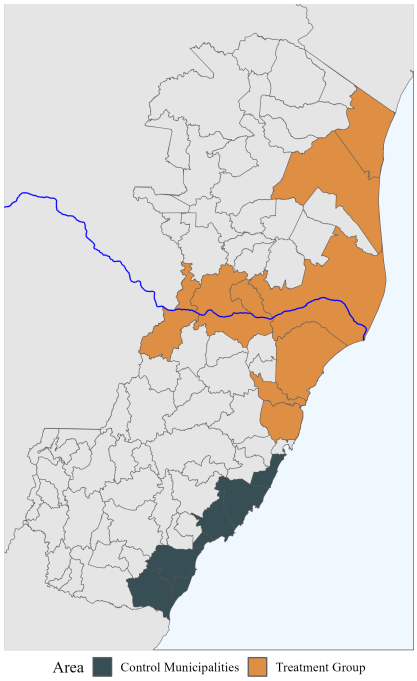
Based on the government's disaster impact reports, which state the contaminated ocean water went northwards towards Bahia, my analysis distinguishes this region into a northern and southern coast. Consequently, the treated units comprise individuals in municipalities crossed by the river or those residing in the northern coastal municipalities identified by the government as having contaminated seawater, as shown in Figure 1.4b.

The control group consists of the southern coastal municipalities unaffected by the pollution and located south of the state capital, Vitória.

My ultimate objective is to extract the effect of the disaster in these regions without the unintended effects of capital disruption of the Iron Square sample. Summary statistics tables for both settings are displayed in the Appendix Section.



(a) Extended Region of Doce River in Minas Gerais.



(b) Espírito Santo State and Affected Regions Near the Atlantic Ocean.

Figure 1.4: Additional regions analyzed. The darkened areas are control municipalities.

Extended Regions Results

Table 1.3 presents the results using log-wage as the outcome variable. The first row shows the results for the aggregate market in the extended region of Minas Gerais. The four columns stand for the same models used in the previous analyses. In the extended Minas Gerais sample, I find that the Mariana disaster significantly negatively impacted wages, as indicated across all models, ranging from approximately 3.8 to 5.8 percent. This trend is followed, to some degree, by the rural activities sector, shown in the second row. However, the effects lose significance when accounting for the balance between control and treatment groups, which means most of these effects, at least for this specific sector, appear due to a lack of counterfactual among the two tested groups. Still, the sign of all estimates is negative, in a similar magnitude range from the aggregate market.

Even though the rural market shows weak evidence of an effect, the same directions of both aggregate and rural markets suggest the environment, represented here by Doce River, plays much more the role of a productivity component in the labor market than an amenity, per Mariana's main results.

Moving onto the Espírito Santo sample, shown in the second part of Table 1.3, while the aggregate effect on log wages presented in the third estimate row is negligible for all models, the sector-specific analysis reveals a different picture. The rural activities in the treated municipalities from Espírito Santo suffered significant negative wage effects across all models, with at least ten percent of statistical significance. When I focus on fishing activity firms only, as shown in the sixth estimate row, the magnitude increases, ranging from around 6.6 and 7.0 percent decrease for the basic fixed effects models to 11.3 and 12.5 for the doubly robust models. Not surprisingly, these wage results suggest that after the river and now the ocean were spoiled, the market's production function had a negative capital shock, pushing wages downwards.

I also further investigate the extended region by providing estimates when using the dismissal and the moving variable in Table 1.4. The first four columns (1-4) are for the linear probability of being dismissed from one's current job, while the last set (5-8) shows the linear probability of moving out from a municipality results. Accordingly, I explore the two additional regions, the extended Minas Gerais and the Espírito Santo coastal region.

For dismissals, the vast majority of the results are insignificant, except for the extended Minas Gerais aggregate market, where it shows a positive change in dismissal probability ranging from 1.9 to 2.3 percent. Even though the magnitude and direction of the estimates can be explained by the productivity shock hypothesis, they do not survive my most robust model of fixed effect interaction and inverse propensity score weighting in Column (4).

Table 1.3: Extended Regions Results

	Log Wage			
	Fixed Effects Models		Doubly Robust Models	
	(1)	(2)	(3)	(4)
Extended Minas Gerais Sample				
Aggregate: Treat x Post	−0.038***	−0.035**	−0.058***	−0.056**
	(0.014)	(0.015)	(0.021)	(0.023)
Rural Activities: Treat x Post	−0.061***	−0.045***	−0.047*	−0.038
	(0.012)	(0.010)	(0.027)	(0.029)
N Clusters	323	323	323	323
N	19 334 590	19 334 590	19 334 590	19 334 590
Espírito Santo Sample				
Aggregate: Treat x Post	0.013	0.003	0.015	0.003
	(0.009)	(0.012)	(0.013)	(0.017)
Rural Activities: Treat x Post	−0.035***	−0.027**	−0.033***	−0.024*
	(0.008)	(0.010)	(0.010)	(0.012)
Fishing: Treat x Post	−0.066***	−0.070***	−0.113***	−0.125***
	(0.010)	(0.009)	(0.019)	(0.016)
N Clusters	16	16	16	16
N	5 100 644	5 100 644	5 100 644	5 100 644
Individual FE	X	X	X	X
Municipality FE		X		X
Individual x Municipality FE		X		X
Year FE	X	X	X	X

[1] Standard-errors are clustered by municipality.
[2] * p < 0.1, ** p < 0.05, *** p < 0.01

[1] Standard-errors are clustered by municipality.
[2] * p < 0.1, ** p < 0.05, *** p < 0.01
[3] Covariates used for propensity score estimation in DR models are age, tenure, education level, gender, race, worker's occupation, and the firm's main economic activity.
[4] Rural Activities correspond to agriculture (01), forestry (02), and fishing (03).

In a similar fashion, the rural activity analysis for the extended Minas area yielded a negative 7.9 percent chance, statistically significant at 5 percent, when using a doubly robust procedure. However, the effect disappears when accounting for interactions between individual and municipality fixed effects. For Espírito Santo, no estimates yielded statistically significant results for the aggregate market and for rural activities. For rural activities, results are negligibly close to zero in magnitude.

Casting our attention to the probability of moving analysis for the extended Minas area, results are inconclusive, with the aggregate market and rural activi-

ties both displaying meaningful effects in Column (7) but not surviving interactions between individuals and municipalities, which suggests the calculated causal effects are driven by changes in the treatment and control composition when individuals move from one place to another.

In contrast, in the Espírito Santo sample, I find a positive effect on moving out probabilities, even with inconclusive results in the dismissal analysis. One explanation could be that the moving-out probability appears conditional on already dismissed individuals. In other words, a worker's position in the fishing firm may not be affected by the disaster, only through wages. However, in the case of a dismissal, this individual is more likely to find another job in a firm outside of the original municipality.

The adverse effects on job security and population stability appear to be less meaningful than the direct productivity losses from capital destruction. Still, it points towards a critical interaction between the environment (as a form of capital component) and local economic conditions. In other words, in a disaster like Mariana's, where the environment is severely spoiled, but human capital is preserved, which is a reflection of drastic environmental changes potentially already occurring due to massive human intervention everywhere, it should be expected an overall impoverishment of the market instead of a search for reallocation.

Nevertheless, given the size and complexity of the studied markets, there are potentially ignored underlying effects when observing specific occupations or using other dimensions in the regressions. Even though, on the aggregate, it seems individuals absorb the shock through wages, there is still the possibility of "Rosen-Roback" effects when accounting for specific demographics or types of firms.

Table 1.4: Extended Regions Results - Probability of Dismissals and Moving Out

	Prob. of Dismissal				Prob. of Moving			
	Fixed Effects Models		Doubly Robust Models		Fixed Effects Models		Doubly Robust Models	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Extended Minas Gerais Sample								
Aggregate:Treat x Post	0.019**	0.023**	0.019**	0.020	0.011*	0.010	0.015***	0.008
	(0.008)	(0.009)	(0.009)	(0.012)	(0.006)	(0.007)	(0.005)	(0.006)
Rural Activities: Treat x Post	−0.011	0.027	−0.079**	−0.004	−0.026	0.009	−0.055*	0.010
	(0.029)	(0.023)	(0.040)	(0.054)	(0.019)	(0.014)	(0.028)	(0.031)
N Clusters	323	323	323	323	323	323	323	323
N	19 334 590	19 334 590	19 334 590	19 334 590	16 541 725	16 541 725	16 541 725	16 541 725
Espírito Santo Sample								
Aggregate: Treat x Post	−0.021	−0.023	−0.035	−0.024	−0.018	−0.016	−0.035	−0.017
	(0.031)	(0.015)	(0.026)	(0.016)	(0.040)	(0.010)	(0.035)	(0.012)
Rural Activities: Treat x Post	0.008	0.010	0.006	0.009	−0.004	0.009	−0.005	0.009
	(0.024)	(0.028)	(0.024)	(0.028)	(0.020)	(0.026)	(0.020)	(0.026)
Fishing: Treat x Post	0.006	0.021	−0.105	−0.074	0.076**	0.094***	0.049	0.101**
	(0.029)	(0.033)	(0.063)	(0.083)	(0.034)	(0.017)	(0.057)	(0.036)
N Clusters	16	16	16	16	16	16	16	16
N	5 100 644	5 100 644	5 100 644	5 100 644	4 351 754	4 351 754	4 351 754	4 351 754
Individual FE	X	X	X	X	X	X	X	X
Municipality FE		X		X		X		X
Individual x Municipality FE		X		X		X		X
Year FE	X	X	X	X	X	X	X	X

[1] Standard-errors are clustered by municipality.

[2] * p < 0.1, ** p < 0.05, *** p < 0.01

[1] Standard-errors are clustered by municipality.

[2] * p < 0.1, ** p < 0.05, *** p < 0.01

[3] Covariates used for propensity score estimation in DR models are age, tenure, education level, gender, race, worker's occupation, and the firm's main economic activity.

[4] Rural Activities correspond to agriculture (01), forestry (02), and fishing (03).

1.7.3 Office Occupations Analysis

Even though the measurements indicate that the effects surfacing on the aggregate market are related to capital shocks, there is a possibility that other heterogeneous effects are at play, which I did not capture in my previous regressions. To understand better the Mariana disaster's implications on the equilibrium relationship between wages and amenities, I perform an occupation-based analysis of workers in the sampled areas.

In accordance with the theoretical framework shown in Section A.1, I choose occupations as the key variable for the analysis due to, in this case, the individual making the decision to stay or move from the affected regions. The challenge of using a firm's economic activities is that, within the same firm, there will be too heterogeneous cohorts of workers at different education levels, types of occupations, and tenure duration. Therefore, a more correct approach is to perform the study taking into account the worker's occupation types across firms.

It is possible to identify the type of occupation through RAIS. I use Brazil's Ministry of Labor's Brazilian Classification of Occupations, CBO in Portuguese, to classify the worker-wise type of activity. The code can be granulated up until six digits, each subsequent digit representing a more specific set of tasks. For instance, the code "21" represents STEM professionals, while "214" represents engineers and architects. A final specification example for this code sequence is "2142-10", representing airport engineers.

Potentially, there are two main types of occupations. Occupations directly and indirectly related to water as a component of labor. For example, farmers primarily rely on water to perform their activities and ultimately provide labor output. On the other hand, office workers, such as accountants and clerks, may only take the water contamination as an amenity nuisance if, for the sake of simplification, we ignore the costs of acquiring freshwater.

This class of occupations [2], which is less directly linked to the local environment in its day-to-day operations compared to the rural activities and fishing sectors, offers a suitable case study of potential Rosen-Roback effects.

Office occupations are primarily influenced by factors such as technology availability, human capital, and infrastructure rather than local natural resources. Therefore, we would typically expect this sector to be less sensitive to environmental disruptions. However, if the Rosen-Roback framework holds, we would anticipate some level of impact from the Mariana disaster, primarily through indirect channels such as reduced local consumption or changes in local labor supply dynamics. Moreover, firms may increase their wages specifically for these

[2] The codes I use for the regressions are 410, 411, 412, 413, 414, and 415. These represent clerks, secretaries, and office administrators, generally responsible for the day-to-day routine operations of offices but not necessarily high-specialized functions such as accountants, lawyers, or bankers.

occupations to compensate for the negative shock on surrounding amenities, which is the key measurement of my study.

My empirical framework, structured similarly to previous sections, utilizes occupation code dummies as interaction terms with the treatment variable. The four columns in Table 1.5 represent the two baseline models and two doubly robust models. Each pair includes an individual fixed effect term and an interaction term between municipality and individual identifiers.

The analysis is carried out across the three key regions, each yielding distinctive insights about the labor market outcomes post-disaster.

The first region of focus is the Iron Square region, where proper Mariana is located and where the disaster took place. Here, the data indicate a shift in spatial equilibrium, as office workers' wages show an increase of approximately 5 percent in the more robust models. Further evidence supporting this equilibrium shift is the 3-6 percent decrease in the likelihood of office workers moving or being dismissed from their jobs when compared to unaffected municipalities in the region.

The subsequent region, the extended region of Minas, displays distinct patterns. While wage change remains neutral, we observe a one percent increase in the likelihood of dismissal. This trend is consistent across all model specifications and aligns with a similar pattern in the probability of moving.

Lastly, the Espírito Santo state analysis offers weak evidence with a slight two percent wage increase at the ten percent significance level. The changes in the likelihood of dismissal and migration out of the region are negligible and statistically insignificant when compared with the control groups.

Despite the variations across regions, these results provide compelling evidence that spatial equilibrium dynamics significantly shape labor market outcomes after environmental shocks, at least for specific categories indirectly related to the environment per se. However, it is crucial to understand these findings in context. As established in the main results, the aggregate market tends to respond to such shocks as capital shocks above all else.

Table 1.5: Office Occupation Analysis

	Fixed Effects Models		DR Models	
	(1)	(2)	(3)	(4)
Mariana Municipality Region				
Log Wage: Treat x Post	0.014	0.020	0.052**	0.051**
	(0.018)	(0.011)	(0.023)	(0.023)
Dismissal: Treat x Post	−0.042	−0.048**	−0.037**	−0.054***
	(0.029)	(0.018)	(0.014)	(0.010)
Mover: Treat x Post	−0.059	−0.062**	−0.033**	−0.052***
	(0.042)	(0.021)	(0.013)	(0.014)
N Clusters	12	12	12	12
N	1 567 721	1 567 721	1 567 721	1 567 721
N (mover)	1 348 847	1 348 847	1 348 847	1 348 847
Extended Minas Gerais Region				
Log Wage: Treat x Post	0.001	0.004	0.008	0.013
	(0.011)	(0.010)	(0.010)	(0.009)
Dismissal: Treat x Post	0.012***	0.012	0.012***	0.011
	(0.003)	(0.008)	(0.003)	(0.008)
Mover: Treat x Post	0.014***	0.012**	0.012***	0.011**
	(0.004)	(0.006)	(0.003)	(0.005)
N Clusters	323	323	323	323
N	19 334 590	19 334 590	19 334 590	19 334 590
N (mover)	16 541 725	16 541 725	16 541 725	16 541 725
Espírito Santo Region				
Log Wage: Treat x Post	0.023*	0.022*	0.025*	0.022
	(0.012)	(0.012)	(0.013)	(0.014)
Dismissal: Log Wage: Treat x Post	−0.012	−0.017*	−0.021	−0.018
	(0.016)	(0.009)	(0.023)	(0.012)
Mover: Treat x Post	0.007	0.002	−0.004	0.000
	(0.017)	(0.006)	(0.026)	(0.007)
N Clusters	16	16	16	16
N	5 100 644	5 100 644	5 100 644	5 100 644
N (mover)	4 351 754	4 351 754	4 351 754	4 351 754
Individual FE	X	X	X	X
Individual x Municipality FE		X		X
Municipality FE		X		X
Year FE	X	X	X	X

[1] Standard-errors are clustered by municipality.

[2] * p < 0.1, ** p < 0.05, *** p < 0.01

[3] Covariates used for propensity score estimation in DR models are age, tenure, education level, gender, race, worker's occupation, and the firm's main economic activity.

[4] The codes I use for the regressions are 410, 411, 412, 413, 414, and 415.

1.8 Conclusion

This study offers a comprehensive analysis of the Mariana disaster's impact on labor market outcomes, juxtaposing two theoretical models in economics: the productivity factor model and the Rosen-Roback spatial equilibrium model. The former posits that the disaster, through water contamination, negatively shocks physical capital levels, thereby driving down wage levels. Conversely, the latter model suggests that the river, as an amenity, should prompt an increase in wages to maintain workers' utility levels in the wake of its destruction; otherwise, workers may relocate to more desirable areas.

Employing a rich administrative dataset encompassing all formal workers and utilizing a difference-in-differences design, the findings indicate significant negative effects on wages in two out of the three studied regions. This suggests that the river primarily functions as capital in the market's aggregate production function. These results persist across alternative fixed effect specifications and "doubly" robust regressions using inverse propensity score weights for estimate calculation.

The results of examining heterogeneous effects across specific industries align with economic intuition. For sectors such as mining, agriculture, and fisheries, where water is integral to operations, the impact on workers' wages was also negative, with varying degrees of robustness. This indicates that the aggregate results mirror underlying disruptive capital shocks closely tied to these economic activities. However, the linear probability models for job movement and dismissal in the aggregate and aforementioned specific markets did not yield strong enough results to establish a conclusive scenario.

Interestingly, when focusing on individuals working in occupations not directly related to water, specifically office workers such as clerks and secretaries, positive pressure on wages and a higher propensity to leave the affected area were observed, aligning with the urban spatial equilibrium proposal.

These findings suggest that, on aggregate, individuals experiencing drastic environmental changes with negligible alterations in human capital, such as potential scenarios of climate change effects, may absorb the shock through the market's production function, leading to wage losses and overall impoverishment. However, the effects of spatial equilibrium should not be underestimated. Policymakers must consider a given region's workforce composition and key characteristics when formulating responses to such environmental disasters.

There is considerable room for more studies regarding the Mariana Disaster. Due to its massive disruption and the complexity of the region, the labor market itself represents one of many aspects potentially affected by the incident. For instance, when analyzing spatial equilibrium, one should also consider the

influence of housing costs and the real estate market. However, I was unable to acquire related data.

In conclusion, this study provides an effort to understand the relationship between environmental disasters, labor market outcomes, and spatial equilibria, highlighting the need for further research to disentangle these effects and inform policy decisions. As climate change continues to pose significant threats to our environment, understanding these dynamics will be crucial for mitigating the economic impacts and fostering resilience in affected communities.

Chapter 2

Labor Market Effects of the Venezuelan Refugee Crisis in Brazil

2.1 Introduction

Forced displacement has become an escalating global crisis, with an unprecedented number of individuals compelled to flee their homes in search of safety and stability. According to the UNHCR, 2023, at the end of July 2023, there were a staggering 110 million people forcibly displaced worldwide, among whom more than 36 million were classified as refugees. Significant attention has therefore been drawn to the potential impacts of refugees on host communities. Extensive research on the labor market effects of refugees largely centers on developed nations, leaving a critical gap in understanding the effects in developing regions, where data scarcity exacerbates the challenge.

We focus on the Venezuelan humanitarian and migration crisis that started in the early 2010s. The drop in global oil prices, Venezuela's largest export, coupled with government mismanagement led to hyperinflation, a shortage of basic commodities such as food and medicine, and an upsurge in crime. As a result, more than five million Venezuelans were forced to flee the country. UNHCR estimated that by 2022, more than four million Venezuelans lived as refugees, mostly in neighboring South American countries. Brazil, in particular, had received approximately a quarter million Venezuelan refugees (UNHCR, 2022), who crossed the border every year exponentially since 2013. Virtually all Venezuelans entered Brazil by land, using the only highway that connects the two countries through the Brazilian state of Roraima.

This study seeks to investigate the causal relationship between the Venezuelan refugee crisis and the labor market in Roraima, the Brazilian border state that directly experienced the influx of Venezuelans. Roraima's geographic isolation from the rest of Brazil provides a unique natural experiment. Using comprehensive administrative panel data on the universe of Brazilian formal workers and utilizing a difference-in-differences approach, we analyze the labor market effects of the refugee crisis in the Roraima labor market in comparison to similar states in north Brazil unaffected by the crisis.

Our main findings show a small but significant positive impact, of around 2 percent, on the average monthly wages of formal Brazilians who lived in Roraima. We show that this increase was not driven by an exit of low-wage Brazilian formal workers from the market as we find no significant effects on job displacement. We show, however, that this effect was primarily driven by complementary dynamics between the formal and informal sectors, and, to a lesser extent, by the fact that the presence of immigrants in the formal market allowed natives to move occupations experiencing a higher wage increase.

Nevertheless, workers in industries and occupations with a higher share of refugees did not experience any significant change in wages during the crisis, potentially due to the substitution effect of immigrants working in the same industry-occupation cells. These findings suggest that when not directly competing with native workers, Venezuelans acted as complements, increasing formal labor market wages. We also conduct analysis using nationally representative survey data and show a pronounced negative wage impact in the informal sector, more evident for individuals involved in industries with a higher immigrant involvement observed in our main administrative dataset. These results are in line with Peri and Sparber, 2009; Manacorda, Manning, and Wadsworth, 2012; Dustmann, Frattini, and Preston, 2013; Foged and Peri, 2016, who argue that immigrants are imperfect substitutes for natives, have a potentially different skill set, and specialize in positive efficiency.

Refugees comprise a distinct subset of immigrants due to their more vulnerable societal position.[3] Unlike other immigrant groups, their displacement is mainly involuntary. A majority of the economic studies on the labor market impacts of refugees focus on developed countries in Europe and North America. These studies revolve around refugee shocks due to a political crisis and utilize natural experiments to analyze the labor market effects of the shock, an approach similar to ours. However, the results from these studies have little overall consensus. While some studies find no adverse effects on native employment and wages (e.g., Card, 1990, Hunt, 1992, Friedberg, 2001), others find large adverse effects (e.g., Glitz, 2012, Dustmann, Fasani, et al., 2017).

[3] In the period of our study, the Venezuelans displaced by the crisis were virtually the only non-Brazilian population in the state of Roraima. Therefore, specifically for this paper, we refer to Venezuelan refugees as immigrants or foreigners, interchangeably.

A small but growing subset of economic literature explores this question from the developing country perspective (Maystadt and Verwimp, 2014; Calderón-Mejía and Ibáñez, 2016; Ruiz and Vargas-Silva, 2016; Taylor, et al., 2016; Alix-Garcia, et al., 2018; Maystadt and Duranton, 2019). The more recent wave of this literature revolves around the Syrian and Latin American crises. Studies on the Syrian refugee crisis have yielded varied findings that range from no effects on employment and wages Fallah, Krafft, and Wahba, 2019 to moderate employment losses among natives Tumen, 2016. Results for Aksu, Erzan, and Kırdar, 2022 are the most similar to ours in that the study finds adverse effects on competing natives in the informal market but positive employment and wage effects for complementary workers in the formal market.

The literature on the labor market effects of Venezuelan immigrants in host countries focuses mostly on Colombia, and, to a smaller extent, on Peru and Ecuador. In Colombia, various studies reveal adverse effects on native employment within the formal labor market and reductions in wages and income in the informal sector (Caruso, Canon, and Mueller, 2021; Delgado-Prieto, n.d.; Lebow, 2022).

Conversely, the labor market analysis in Peru presents a contrasting perspective. In line with our results, Groeger, León-Ciliotta, and Stillman, 2024 indicates that the influx of Venezuelans to specific locations in Peru has resulted in an upswing in both employment and income among the local population. Morales and Pierola, 2020 finds that an increase in the share of Venezuelan migrants in Peru is associated with an increase in the probability of being employed for Peruvians in the non-service sector although accompanied by a decrease in the probability of having an informal job.

Bahar, Ibáñez, and Rozo, 2021 studies the labor market impacts of an extensive migratory amnesty program that granted work permits to nearly half a million undocumented Venezuelan migrants in Colombia in 2018. Their analysis indicates no significant impact of the program on hours worked, wages, or labor force participation of Colombian workers. Nonetheless, they study only the immediate effects of the amnesty program without considering its potential dynamic effects in the years following the treatment.

Ryu and Paudel, 2022 addresses this question on the labor market effects of Venezuelan immigrants from the Brazilian perspective. Using a national quarterly household survey, Ryu and Paudel, 2022 uses the synthetic control method to study the labor market impacts of the Venezuelan refugee crisis in Roraima, the affected state. The study finds that the crisis lowered labor force participation and employment rate in Roraima and did not find any effects on

wages. However, the dataset did not distinguish Brazilians and Venezuelans, which can underestimate the effects of refugees.

We build upon these studies by employing a rigorous administrative panel dataset of the universe of formal workers in Brazil that has two crucial advantages over Ryu and Paudel, 2022. First, our data allows us to distinguish the nationality of individuals and isolate our sample to Brazilians. Second, because of its panel nature, it allows us to track individuals over time, irrespective of their mobility between states.

We organize the remainder of the paper as follows. Section 3.2 provides the background. Section 2.3 discusses our identification strategy and empirical methodology. Section 3.3 describes the data. Section 21 presents the results and robustness checks of our model. Section 2.6 discusses the mechanisms through which immigrants affect native wages before Section 3.7 concludes.

2.2 Background

This section is divided into two parts. In the first part, we briefly overview the Venezuelan crisis. In the second, we explain the interaction between Venezuelan refugees and the Brazilian labor market, highlighting the distinctions between formality and informality.

2.2.1 The Venezuelan Political and Refugee Crisis

Until the early 2000s, Venezuela had one of the highest GDP per capita in Latin America. With what is considered the largest oil reserves in the world, their GDP was tied to oil exports (EIA, 2019; Haider, 2020). However, the drop in oil prices in the early 2010s severely hit the Venezuelan economy. Coupled with government mismanagement, it led to an unprecedented humanitarian crisis in the country with hyperinflation, a shortage of basic goods and services such as food and medicine, and a rise in crime. Although Venezuelans had already started to leave for other countries in 2011 as a result of the political crisis, the exodus skyrocketed in 2013. The United Nations High Commissioner for Refugees (UNHCR) estimated that by 2022, more than four million Venezuelans lived as refugees, mostly in neighboring South American countries. At the end of 2021, Brazil hosted 260,000 Venezuelan refugees. Due to their geographical proximity, most of the refugees in Brazil were concentrated in the state of Roraima. The Brazilian Federal Police border patrol reported that the number of Venezuelans who entered Roraima in 2017 and stayed in Brazil numbered more than 50 thousand (Lopes, 2018). This number corresponds to 8 percent of the total population of Roraima.

In 2019, the International Migration Organization (IOM) and the Migration Policy Institute (MPI) surveyed thousands of Venezuelan refugees living in 11 Latin American and Caribbean countries (Echeverría-Estrada, 2020). The report found that among the Venezuelan refugees in Brazil, almost 60 percent of the refugees were women, 55 percent were between 25-45 years, and 60 percent had secondary education. The entirety of the sampled individuals had come to Brazil by land, entering through the state of Roraima. Before migrating to Brazil, 65 percent were employed, of which 32 percent were self-employed. 26 percent were unemployed and only 5 percent were students. After immigration, only 40 percent were employed in 2019, of which 29 percent were self-employed, and 58 percent were unemployed.

Due to the only land border crossing between Venezuela and Brazil located in the Brazilian state of Roraima, the state hosted most Venezuelan refugees. Nevertheless, the insufficient public infrastructure and limited job opportunities in Roraima put a strain on the state's health, safety, and education systems (Oliveira, 2019). The increasing number of refugees prompted the federal government to intervene and temporarily suspend the state's autonomy within the federation in December 2018.

The Brazilian federal government adopted a notably generous stance towards accommodating Venezuelan migrants. For instance, it established nine shelters, eight in Boa Vista. Additionally, the Brazilian government enabled Venezuelan migrants to seek employment by granting them work permits, allowing them to function as regular employees in Brazil for up to two years under temporary residency (Ramsey and Sánchez-Garzoli, 2018; Ryu and Paudel, 2022). Furthermore, there are policies aimed at dispersing migrants from Roraima to other states, such as Rio de Janeiro or São Paulo, starting in 2018, although this effort did not significantly impact the relocation of migrants during the periods examined in this paper.

2.2.2 Brazilian Labor Market

Mercosul (or Mercosur in Spanish) is an economic block comprised of South American countries, including Venezuela and Brazil. Members of the block are entitled to free entry, residency rights, and the ability to work in the host country's formal labor market, subject to government authorization. Recently, the Brazilian government has offered Venezuelan refugees a special status that accelerates their permission to work in the formal labor market. Venezuelan refugees must undergo a specific process and submit certain paperwork to obtain this status, which takes at least several months.

In Brazil, any organization must have a National Legal Persons Registry number (CNPJ) to operate legally. The entity's owners must declare its purpose and intended activity to the government. If an entity is not registered, it is considered informal. The costs associated with registering a company can be relatively high, leading to a higher prevalence of informality in poorer regions.

Legal firms are only permitted to hire workers formally, a requirement often disregarded in impoverished areas such as Roraima. Formality in the hiring sense means that employees are entitled to social security and certain rights that the employer must guarantee. Generally, workers in the formal market earn more due to these social benefits than their informal counterparts. However, employers may be at risk of labor litigation if they are caught hiring informal workers. For example, informality in northern Brazil, where we base our study, around the period of our research, corresponded to 45 percent of its total labor market (Azeredo, 2019). Time was likely the main cost of entry for Venezuelans to find regular employment. As more immigrants arrived and application lines increased exponentially, we posit that a fraction of these refugees procured economic activities outside formality.

2.3 Methods

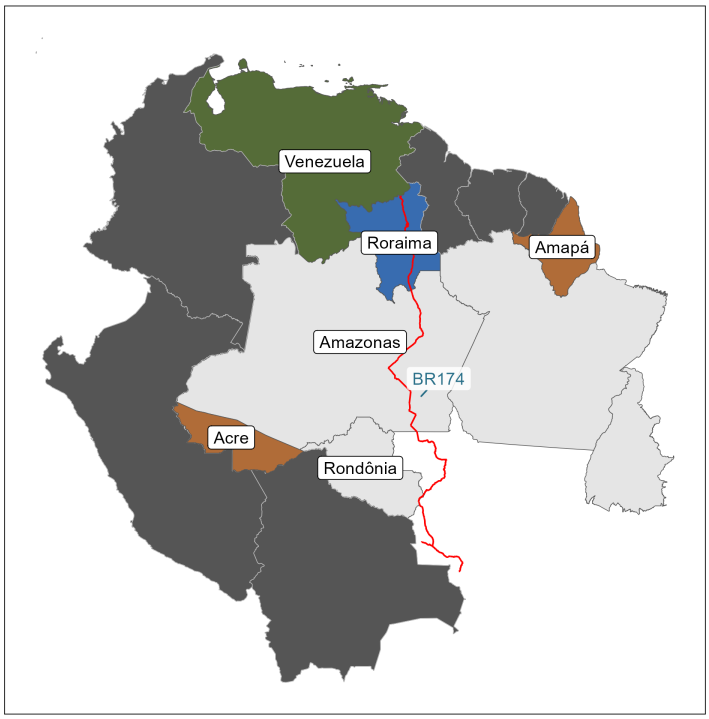
In this section, we discuss the identification strategy and the empirical strategy used to analyze the labor market impacts of the Venezuelan refugee crisis in Roraima, Brazil. The identification strategy section highlights the geographical setting of the natural experiment, while the empirical strategy section explains the econometric models used to estimate the causal effects.

2.3.1 Identification Strategy

Geography is crucial for our identification strategy. As shown in Figure 2.1, Venezuela borders the Brazilian states of Amazonas and Roraima, but the Amazon rainforest makes it highly unfeasible to enter Brazil through Amazonas. Instead, people can enter Brazil from Venezuela through the state of Roraima. The BR-174 highway, represented by the solid vertical line that traverses Roraima and Amazonas in Figure 2.1, is the only land transportation route between the two countries, which runs through Roraima from the Venezuelan border. This makes Roraima a key transit point in Brazil, but once refugees are in Roraima, they can only practically go as far as Amazonas by land. To reach larger coastal cities, they would need to use air routes. Therefore, Roraima is isolated from the rest of the northern states, which leads many refugees to choose to enter and stay there.

This setting provides a natural experiment in which Roraima acts as the treated group. For our control group, we use states that share similar sociogeographical characteristics and are located on the Brazilian border: Acre borders Peru and Bolivia, while Amapá borders Suriname and French Guiana. Like Roraima, these states also have a large portion of their population concentrated in their respective capital cities and are relatively isolated. In particular, Amapá does not have an inland connection to the rest of Brazil and can only be reached by airplane or boat crossing the Amazon hydrographic basin to reach its capital, Macapá.

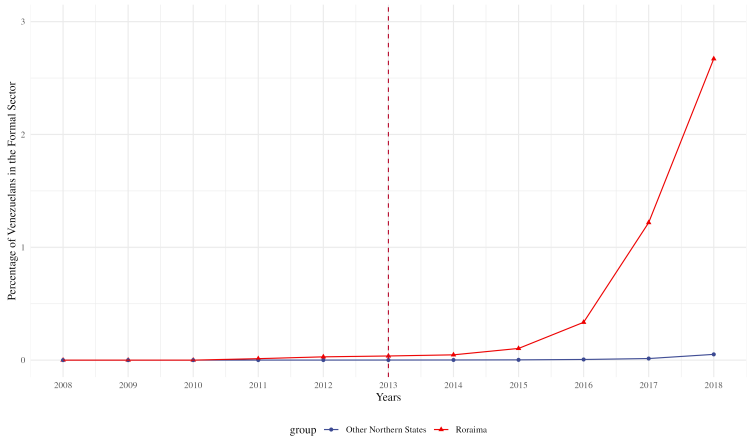
Figure 2.1: Map of Venezuela and Brazil's North Region



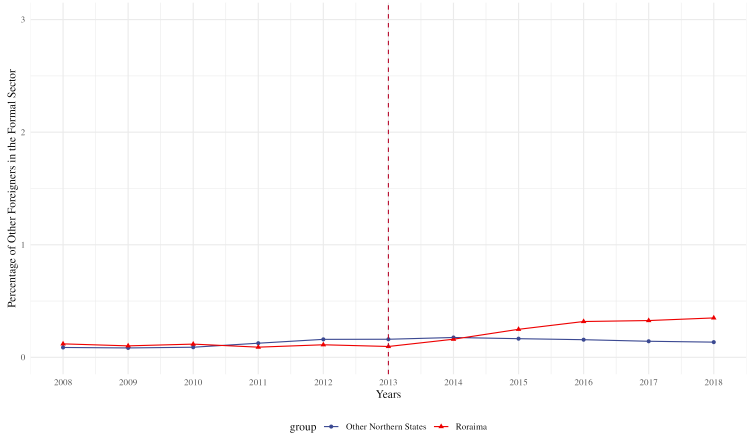
According to Borjas, et al., 1997; Borjas, 2003; Borjas, 2006, it may be difficult to accurately assess the impacts of immigration on the labor market by considering geography alone due to the potential for spillover effects across regions. However, in our case, Roraima is isolated with high transportation costs, which limits mobility between states or even within municipalities in the same state.

Foreign Presence in RAIS

A key assumption of our paper is that refugees were not attracted to high-growth regions, but ended up in Roraima due to its geographical isolation and high transportation costs to move to cities farther from the border. To verify, we can use our administrative data to check the location of Venezuelan refugees in our sample. We can also examine wage trends before the crisis to confirm that Roraima's growth trend was not significantly different from the control regions, which we explore in the next subsection. Figure 2.2a shows the annual percentage of Venezuelans in the labor market, grouped by treatment and control states. In 2017, Venezuelans represented about 2 percent of Roraima's formal labor market, while none were observed in Amapá or Acre. Figure 2.2a supports our assumption by demonstrating that Venezuelans largely remained in Roraima.



(a) Venezuelans as % of Labor Market



(b) Other Foreigners as % of Labor Market

Figure 2.2: Proportion of non-Brazilians in the formal labor market for Roraima and the control states

It would be problematic for our experiment setting if any other nationality besides Venezuelans had a substantial growth in population, either in Roraima or the control states. To ensure that the exponential growth of foreigners we observe is due to Venezuelan refugees, we construct Figure 2.2b by aggregating data on all immigrant nationalities other than Venezuelans and plotting their proportion in the formal labor market. As the figure shows, there is negligible growth in the control states and no significant growth in Roraima following 2014. The non-zero values we observe for non-Venezuelan immigrant nationalities can be attributed to several factors. First, there is a significant presence of Haitian immigrants in control states due to United Nations peace operations in Haiti that began in the 2000s and remained constant over the study years. Second, there are observations of other Latin American and Bolivian nationals

in the data in control states due to the proximity of Acre to Bolivia and the resulting natural population exchange. Finally, some of these foreigners may be Venezuelans with dual citizenship, as they only appeared in Roraima after the crisis began, as Figure 2.2b shows. Since there is marginal change in these populations in either treatment or control groups, the non-zero values should not pose a threat to our identification strategy.

Although the administrative data provides a good overview of Venezuelans entering Brazil and remaining in the state of Roraima, it only includes those in the formal labor market. It could compromise our identification strategy if the control states also had many Venezuelan refugees outside the formal labor market. In Appendix B.1, we use refugee application data to demonstrate that this is not the case. We show that only Roraima, not the control states, experienced a significant increase in the number of Venezuelan citizens seeking refugee status applications.

2.3.2 Empirical Strategy

To empirically assess the effects of the refugee crisis in Roraima labor market, we estimate equation (2.1).

$$y_{imt} = \theta_i + \alpha_t + \beta^{bin} D_{imt} + \epsilon_{imt} \tag{2.1}$$

In this model, y_{imt} denotes the logarithm of the average monthly wage for individual i working in municipality m during year t. The term θ_i captures time-invariant characteristics unique to each worker, which in our dataset are represented by a series of social identifier dummies. The α_t term represents year fixed effects, accounting for broader temporal trends and seasonality that could affect wages across all states.

Our primary focus is on the parameter β^{bin}, which quantifies the returns of being exposed to Venezuelan immigration after 2013. This exposure is encapsulated in the binary treatment variable D_{imt}, which takes a value of 1 for individuals in the municipality that belongs to Roraima post-2013 and zero otherwise. The term ϵ_{imt} denotes the idiosyncratic error term, capturing unobserved influences on wages. Since our observed treatment heterogeneity can occur across regions within each state, we cluster at the municipality level, as discussed in Abadie, Athey, et al., 2023.

To further refine our analysis, we introduce an alternative specification where the treatment effect is modeled as a continuous variable, following Callaway and Sant'Anna, 2021a. This approach uses the ratio of Venezuelan refugees per municipality as a proxy for exposure to the immigration shift, as stated in

equation (2.2). Here, R_{imt} represents the ratio of immigrants per municipality m where individual i is located at a given year t.

$$y_{ist} = \theta_i + \alpha_t + \beta^{cont} R_{imt} + \epsilon_{ist} \tag{2.2}$$

Both approaches exploit the within-individual variation in the exposure to whether the state experienced the refugee crisis. Our identifying assumption is that the time-varying shocks are orthogonal to the treated state.

Doubly-Robust Estimator

To strengthen the credibility of our causal inference, we employ a doubly robust estimation framework, balancing observable characteristics of workers in our data through propensity scores and applying the inverse of these scores as weights in our regressions. This approach ensures that particular features across control and treatment groups will not bias our results. Moreover, it has the advantage that our results are consistently estimated even if one of the difference-in-differences or the propensity score specifications is not (Uysal, 2015; Sant'Anna and Zhao, 2020).

In the first step of the doubly robust framework, we use the following logistic regression:

$$p(X_i) = Pr(Z_i = 1|X_i) = F(\theta' X_i) = \frac{e^{\theta' X_i}}{1 + e^{\theta' X_i}} \tag{2.3}$$

In this regression, $P(Z_i = 1|X_i)$ represents the probability of being in the treatment group given the vector X_i of time-invariant covariates for individual i and Z_i represents treatment group membership, which is working in Roraima. The parameter θ establishes the logistic relationship between the covariates and the group membership.

The second step is to construct the inverted propensity scores based on control and treatment group.

$$w_i = \begin{cases} \frac{1}{1-p(X_i)}, & \text{if } Z_i = 0 \\ \frac{1}{p(X_i)}, & \text{if } Z_i = 1 \end{cases} \tag{2.4}$$

We also impose trimming rules to avoid propensity score values that are too extreme, avoiding weights above the 99.75 percent quantile (Lechner and Strittmatter, 2019). The final step minimizes the weighted sum of squares in the two-way fixed effects framework. For instance, in our binary treatment case,

we have:

$$\hat{\beta}^{dr}_{bin} = \arg\min_{\beta} \sum_{i=1}^{N} w_i (y_{imt} - \beta D_{imt} - \gamma_i - \omega_t)^2 \tag{2.5}$$

Where w_i is calculated using Equation (2.4). N is the number of observations in the sample. The analogous approach is employed for the continuous treatment, in which we replace D_{imt} to R_{imt}.

2.4 Data

Our primary dataset is the Annual Social Information Survey (RAIS) maintained by the Brazilian Ministry of Labor and Employment. RAIS contains the universe of the Brazilian formal labor market, comprising panel information on individual workers and establishments, including employment status, occupation, industry[4], and wages. We focus on observations from 2008 to 2018. Our pre-treatment years are 2008-13, and the post-treatment period includes 2014-18. We use this cutoff since the refugee crisis intensified in 2014 based on the data and journalistic accounts (see the identification strategy section for more details).

[4] Throughout this paper, we use the term 'economic sector', 'industry', or 'economic activity' to refer to the industry in which a firm operates, such as retail, construction, or restaurants. We use the term 'occupation' to describe the specific job title or role of the worker within that sector, such as janitor, accountant, or waiter.

The Brazilian equivalent of a Social Security Number, PIS (Social Integration Program), identifies unique persons and is present in the data. Along with information on the labor market at the individual level, the data also provides sociodemographic information, including nationality. For occupation, we use the first three digits of the Brazilian Occupation Code (CBO) to categorize, while for the economic sector, we use the first 2 digits of the National Registry of Economic Activities (CNAE).

The dataset is constructed based on the contract generated between the employee and the firm. Therefore, the base observation is contract-year. To convert the cells to individual-year observations, we restricted the data to Brazilians in a given year employed in the private sector with more than or equal to 30 weekly hours of labor. If there are multiple observations for the same individual, we use the contract with the oldest tenure and the highest pay (Lavetti and Schmutte, 2023).

In our dataset, payments are recorded as the average monthly compensation each individual receives according to their contract. To construct our primary outcome variable, we exclude individuals with recorded payments of zero. Furthermore, we apply censoring to this variable at both tails of the distribution: at the 99.75th percentile from the upper end and the 0.25th percentile from the lower end. This measure mitigates the influence of extreme outliers. Finally, we

transform the variable into an hourly wage measure by integrating it with the recorded weekly hours worked times four.

2.4.1 Summary Statistics

By restricting our sample to Brazilian workers who were employed in the private sector with at least 30 weekly hours of labor, we retain the sample size of 58,379 individuals, of whom 12,581 are from Roraima and 45,798 from control states. Table 2.1 summarizes the socioeconomic and demographic characteristics of Brazilians in our treatment state, Roraima, and our control states in the RAIS datasets in 2013, the year right at the start of the refugee crisis. The average monthly wages for Roraima and control states were comparable at R$ 1,916 and R$ 1,841 respectively. These wages are reported in local currency adjusted to 2018 R$. In 2018, the national minimum wage was around R$ 1,000. The average earnings in these states were, on average, twice this value. Other variables such as average age, experience, fraction of female workers, racial composition, and workers by economic sector and occupation type were all comparable between Roraima and the control states.

Table 2.1: Descriptive Statistics of Natives from Roraima and Control States

	(1) Roraima	(2) Control States
Mean monthly wage	1916.07	1841.05
Mean age	36.92	37.85
Mean experience	60.75	63.70
Fraction of females	0.36	0.32
Race by fraction:		
White	0.22	0.17
Black	0.02	0.03
Mixed	0.63	0.65
Color not declared	0.12	0.15
Education by fraction:		
High school dropouts	0.24	0.32
High school graduates	0.64	0.58
College graduates	0.12	0.10
Economic activities by fraction:		
Commerce	0.41	0.38
Construction	0.08	0.09
Extraction industries	0.03	0.05
Hotels and restaurants	0.04	0.03
Manufacturing and utilities	0.10	0.10
Other services	0.35	0.35
Job occupations by fraction:		
High school level technicians	0.07	0.08
Factory occupations	0.22	0.27
Services	0.43	0.41
Retail and wholesale	0.14	0.11
Rural occupations	0.02	0.03
Scientific and liberal arts	0.07	0.05
N	12 581	45 798

Note: Observations represent full-time workers working in 2013 for both Roraima and the Control States (Amapá and Acre). Mean experience in months. Economic activities are classified by the first two digits from their respective code. Occupations are likewise classified by the first three digits.

Figures B.6 and B.7 illustrate Venezuelan formal workers by economic sectors and occupations respectively in Roraima using 2018 RAIS data, the final year in our data and the year with the largest number of Venezuelans observed in the data. Figure B.6 shows that most of the Venezuelans in Roraima were in the retail commerce sector, followed by restaurants, construction, wholesale commerce, and gardening and landscaping. Figure B.7 shows that most of them worked as general service employees, salespersons, machine operators, and office clerks, which are semi-skilled or low-skilled.

Table 2.2 compares Brazilians with Venezuelans in the formal labor market in Roraima in 2018, the year with the most Venezuelans in our sample period. While the reported race is comparable between the two groups, a much higher percentage of Venezuelans reported to be of mixed race. While men and women were almost equally represented among Brazilians in the Roraima formal labor market, it was much more dominated by men for Venezuelans, at around 76 percent. This comes in stark contrast to the fact that around 60 percent of the Venezuelan refugee population in Roraima was female (Echeverría-Estrada, 2020), highlighting that male Venezuelans were more active in the formal labor market.

Venezuelans tend to have a higher proportion of individuals with only high school education and a markedly lower percentage of college graduates compared to their Brazilian counterparts. Table 2.2 also highlights the top five economic activities and occupations with the highest Venezuelan representation. This was determined by calculating the ratio of Venezuelans to total workers in each category, using the top 2 digits of both to allow for a broader categorization. Pictures B.6 and B.7 were generated using the same procedure.

The educational profiles of the two groups are mirrored in the sectors they predominantly engage in. Immigrants are primarily employed in industries like commerce, restaurants, and construction, which typically require less specialized education. They are mostly found in industries such as restaurants, retail, and construction, encompassing auxiliary jobs across these various industries through the "General Service" category in occupation. In contrast, Brazilians exhibit a more diverse industrial distribution, notably in sectors we categorize as "Other Industries", which includes professions in teaching, banking, legal services, healthcare, and more. Additionally, a significant portion of Brazilians occupy mid-level office positions, classified as 'Office Clerks'.

Finally, the data also reveal a stark contrast in earnings between natives and immigrants. On average, Venezuelans are younger and earn significantly lower wages than Brazilians. This wage disparity may reflect differences in job types, experience levels, or the indirect costs of immigration.

Table 2.2: Statistics of Natives and Venezuelans in 2018

Percentage of	(1) Brazilian	(2) Venezuelan
Race		
White	6.18	5.85
Black	1.07	1.76
Indigenous	0.33	0.03
Mixed	34.04	52.94
Not Declared	58.38	39.43
Gender		
Female	48.68	23.96
Male	51.32	76.04
Education		
No High School	16.97	16.29
High School	54.97	76.02
College	28.05	7.69
Industry by Economic Activities[1]		
Retail Commerce	16.98	38.01
Restaurants	2.74	15.90
Construction	2.56	6.59
Wholesale Commerce	3.11	5.68
Gardening and Landscaping	3.59	1.70
Other Industries	71.03	32.13
Occupations[1]		
General Service Employees	17.37	34.69
Retail and Wholesale Salesperson	8.65	14.42
General Construction Workers	2.93	10.76
Machine Operators	4.23	6.19
Office Clerks	28.80	5.59
Other Occupations	38.03	28.36
Mean Age	36.86	31.67
Mean Wage	2857.94	1198.41
Total Observations	128 389	3523

[1] See text. Occupation and industry identifiers from the data are provided by the Registry of Brazilian Occupations and Registry of Economic Activities. For both variables, we use the code top 2 digits to generate the tables. However, for our regressions, we allow occupations to be detailed by their top 3 digits.

2.5 Results

In this section, we present the main results of our analysis. In the first set of regressions, we explore the effects of Venezuelan immigration on the logarithmic average monthly wage of natives using a difference-in-differences strategy.

2.5.1 Main Results

Columns (1) and (2) in Table 2.3 report our coefficients of interest, β^{bin} and β^{cont} respectively, representing the returns on wages by being in Roraima after the Venezuelan crisis. On average, Roraima experienced a slight positive wage increase after the influx. The interpretation of the continuous approach is that for every 1 percent increase in the presence of Venezuelans in the formal labor market, the wage increased by 2 percent, also consistent with the 2.2 percent increase in the binary treatment variable regression.

Table 2.3: Main Results - Roraima x Amapá and Acre

	Log Wage		Job Retainment	
	(1)	(2)	(3)	(4)
Treat: Binary	0.022***		−0.010	
	(0.006)		(0.010)	
Treat: VZ Ratio x 100		0.021***		0.001
		(0.003)		(0.005)
Individual FE	X	X	X	X
Year FE	X	X	X	X
R2 Adj.	0.875	0.875	0.150	0.150
RMSE	0.245	0.245	0.370	0.370
N Clusters	53	53	53	53
N	577 552	577 552	577 552	577 552

[1] Standard-errors are clustered by municipality.
[2] Propensity score explanatory variables are gender and race indicators, age, age-squared, tenure, tenure-squared and education level.
[3] * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

To address the possibility that our observed wage effects are influenced by low-wage earners exiting the labor market due to immigrant replacement, we employ a linear probability model in line with equations (2.1) and (2.2). Our

focus is to evaluate the likelihood of a worker departing from their job within a given year, a key concern in the context of immigrant labor dynamics. This aspect of our analysis helps determine whether wage increases could be falsely appearing as a result of lower-paid workers being replaced by immigrants. We identify job departures in our dataset through a specific dummy variable, which indicates whether a worker remained with the same firm until the end of the year.

The results are detailed in columns (3) and (4) of Table 2.3. Notably, we cannot statistically reject the null hypothesis of the immigrant crisis affecting job retention in Roraima. These findings provide strong evidence that the observed positive wage effects are not merely a consequence of lower-wage earners exiting the labor market post-immigration. This stability is the first step to indicate that the wage increase is not artificially inflated by a reduction in the number of low-wage positions but rather reflects a genuine enhancement in wage levels. Such a scenario aligns with economic theories that posit immigration can have a complementary effect on the native workforce, potentially due to increased labor demand and productivity. Moreover, it could also be the complementarity effect of informal and formal sectors due to a larger concentration of immigrants outside the formal market. We will discuss these mechanisms in more detail in the following sections.

Ryu and Paudel, 2022 finds that while labor force participation decreased among working-age individuals in Roraima, there were no significant impacts on hourly wages in Roraima. Their use of the cross-sectional PNAD-C data drives the differences in the results as PNAD-C cannot distinguish the nationality of individuals in the sample, therefore including recently arrived Venezuelan migrants in the analysis, who were either searching for jobs or were involved in transitory low-wage jobs. This inherently overestimates the impacts on labor force participation and underestimates the impacts on wages. Our results come from the longitudinal individual-level data distinguishing Brazilians and non-Brazilians, eliminating this issue, and ultimately illustrating positive wage effects in Roraima and the absence of job loss among formal workers.

The key assumption of our strategy is that in the absence of the treatment, our outcome variable will exhibit identical trends in treatment and control states. While we cannot test the counterfactual in post-treatment years, the pre-treatment trends between the treated and control for the outcome variable should be parallel over time.

To test this assumption of parallel trends, we employ an event study method using the binary treatment variable where β^{bin} is disaggregated for every year present in the data. The reference year for comparison is 2013, the year before

the Venezuelan crisis and the refugee influx gained momentum. The estimation procedure is shown in Equation (2.6), where the indicator function D_{imt} now takes a separate value for each year within the summation. β_t^{dr} explains the average difference between treatment and control groups for each year t compared to the reference year of 2013.

$$y_{imt} = \sum_{\substack{t=2008 \\ t \neq 2013}}^{2018} \beta_t^{dr} D_{imt} + \theta_i + \alpha_t + \epsilon_{imt} \tag{2.6}$$

If control and treated groups are comparable before treatment, then, on average, $\beta_t = 0$ for $t \in \{2008, \ldots, 2012\}$. In other words, there should be no difference in trends between treatment and control groups before the treatment. Assuming the only disturbance in Roraima's job market after 2013 is the immigration flow, any variation in the post-treatment period estimates in Roraima must be associated with the refugee crisis. If the refugee crisis affects wages, then the effects should be increasing over time due to the increase in the total refugee numbers. Figure 2.3 shows the event study estimates, which precisely illustrate this pattern.

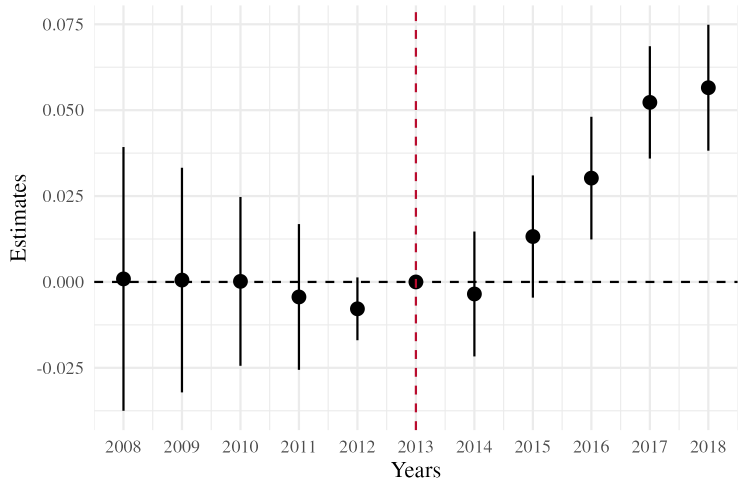


Figure 2.3: Log Monthly Wages Effects Event Study

The results reveal a parallel trend in earlier years right before treatment. They yield pre-treatment estimates not statistically different from zero, with an increasingly upward trend in post-treatment periods. This suggests the difference-in-differences model captured the positive effect of Venezuelans in the formal market, allowing Brazilians, on aggregate, to increase their wages.

2.5.2 Synthetic Control Method

Concerns may arise in our estimation strategy of using the doubly robust difference-in-differences method given the nature of our natural experiment with only one treatment state (Roraima) if the control group is chosen incorrectly. In the main analysis, we compare the treatment state with carefully chosen control states in Brazil with states with socioeconomic and sociodemographic characteristics comparable to those of the treatment state. We also showed pre-treatment parallel trends in our outcome variable. However, it may lead to biased estimates if the control groups are characteristically different from the treatment state.

A popular way of estimating the causal effects of the treatment in the case of a singular or a few treatment states is by using the Synthetic Control Method (SCM). Using SCM with the northern Brazilian states as our donor pool for the synthetic Roraima, we confirm the validity of our choice of control states in the main analysis. We present the details in Section B.2, but from Fig B.3, we observe perfect parallel trends between Roraima and synthetic Roraima until 2013 and a few years following. The trends started to diverge in 2016 where the wages for synthetic Roraima have a reduction in slope but Roraima's slope reduction is less steep. In the year 2018, we see an almost 0.20 percentage point difference between Roraima and synthetic Roraima. This exercise confirms the results from our primary analysis using difference-in-differences. We also provide results from the Synthetic Difference-in-Differences model (SDID) (Arkhangelsky, et al., 2021) in Section B.3 where the outcomes closely mirror our primary results showing an upward wage trend in Roraima compared to synthetic Roraima.

2.6 Mechanisms

In this section, we explore the mechanisms through which Venezuelan immigration might be elevating overall wages in Roraima's formal labor market. We identify three primary mechanisms through which this impact could manifest:

1. Immigrants in the formal labor market can affect wages by increasing worker efficiency and offsetting any substitution effects, resulting in an overall positive change.

2. The informal sector, presumably with a larger concentration of immigrants, can increase overall wages in the formal sector through complementarity.

3. The presence of the immigrant population as a whole, which by 2018 corresponded to 10 percent of Roraima's population, can increase native

wages by boosting consumption and ultimately labor demand. However, we cannot currently test for this with our labor market data.

2.6.1 Effects Due To Presence of Immigrants in the Formal Sector

Economists often consider that immigration is not evenly balanced across groups of workers. For example, if high school graduates are the majority of Venezuelan immigrants, they potentially compete with native high school graduates, but not necessarily with individuals holding a college degree (Card, 2005; Card, 2009; Borjas, 2017; Llull, 2018). Another dimension is occupation, wherein immigrants tend to occupy manual-intensive or low-skilled jobs (Foged and Peri, 2016), which could increase the efficiency of the market and allow the overall average wages to grow. We test for treatment heterogeneity by education and by industry and occupation with immigrant presence.

Treatment Heterogeneity by Education

We now explore potential treatment heterogeneity by education. We categorize our data into three groups: individuals with a college education, individuals with a high school education but no higher degree, and individuals with less than a high school education. Given that most of the observed Venezuelan immigrants in RAIS are high school graduates, we anticipate distinct effects within this cohort.

Table 2.4 presents the results of our difference-in-differences analysis, considering the heterogeneous effects by education level. Concurrently, the corresponding event studies are depicted in Figure 2.4. The two rows of the table delineate the results using the binary treatment variable and the continuous variable respectively. Columns (1) and (2) use a sample of Brazilians with at least a college education. Results indicate approximately a 3.7 percent wage decrease under the binary treatment variable and a 1.3 percent decrease per 1 percent increase in the presence of Venezuelans in Roraima. While college-educated individuals in Roraima appear to be the most adversely affected by immigration, the downward trend in college graduate wages, as illustrated in Figure 2.4c, suggests that this negative impact is compounded by pre-immigration variations.

Columns (3) and (4) use a sample of individuals with a high school education or more. Results show a wage increase of around 1.5-1.7 percent for both types of treatment variables. These figures are markedly lower in magnitude compared to their counterparts with less than high school education, as shown in columns (5) and (6), with wage increases as high as 4.1-4.2 percent. With high

school graduates constituting 70 percent of the Venezuelan population in Roraima's formal labor market, the event studies reveal that these effects emerge post-immigration. This weaker impact for high schoolers might be attributable to the Venezuelan presence in the formal market generating a substitution effect, hampering any external wage growth factors. Yet, the presence of significant positive outcomes suggests that any substitution effects are still likely counterbalanced by other factors. This finding is consistent with Card, 1990 and Clemens and Hunt, 2019, observing no substitution effects on the labor market when analyzing wage level changes due to immigrant influx in lower education individuals. Furthermore, the overall wage increase in the aggregate market implies that immigrants in general primarily serve as a complementary effect more than a substitution effect.

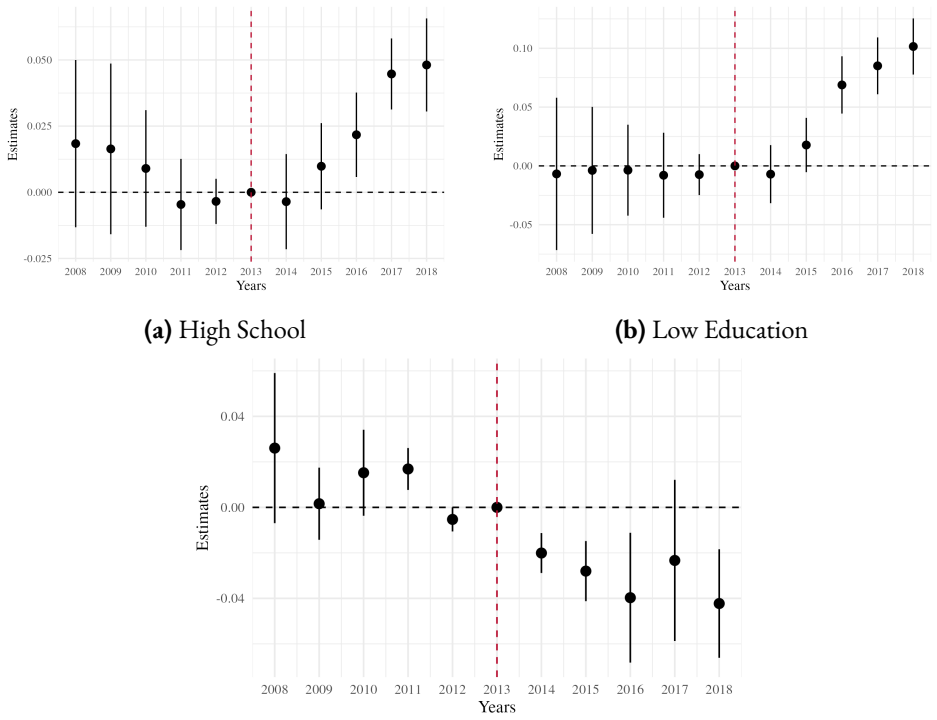
Table 2.4: Heterogeneous Treatment Effects by Education Cohorts

	College Education		High School Education		Low Education	
	(1)	(2)	(3)	(4)	(5)	(6)
Treat: Binary	−0.037***		0.015***		0.041***	
	(0.006)		(0.004)		(0.012)	
Treat: VZ Ratio x 100		−0.013***		0.017***		0.042***
		(0.003)		(0.003)		(0.007)
Individual FE	X	X	X	X	X	X
Year FE	X	X	X	X	X	X
R2 Adj.	0.886	0.886	0.845	0.845	0.857	0.857
RMSE	0.272	0.272	0.236	0.236	0.181	0.181
N Clusters	53.000	53.000	53.000	53.000	53.000	53.000
N	60 575	60 575	335 147	335 147	176 071	176 071

[1] Standard-errors are clustered by municipality.

[2] Propensity score explanatory variables are gender and race indicators, age, age-squared, tenure, and tenure-squared.

[3] * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$



(a) High School

(b) Low Education

(c) College

Figure 2.4: Event Studies for Heterogeneous Treatment Effects by Education Cohorts

Heterogeneity by Industry and Occupation

To investigate further whether immigrants in the formal labor market affected wages, we now focus on the channels of occupation and economic sectors. We create a variable to measure the Venezuelan-to-Brazilian ratio in Roraima across various top 2-digit economic activities in our data. We then divided native workers into two groups: those in sectors with no Venezuelan presence (zero ratio) and those in sectors with Venezuelan presence (positive ratio).

Our preliminary observations indicated that immigrants in RAIS were primarily employed in sectors requiring manual labor, such as retail and wholesale, restaurants, construction, and gardening. Conversely, sectors like finance, insurance, telecommunications, research, pharmaceutics, and entertainment saw little immigrant participation. We applied a similar methodology for occupation analysis, using more detailed 3-digit codes from our dataset. The results of these analyses are presented in Table 2.5.

Columns (1) and (2) of the table display the outcomes for economic activities and occupations with positive immigrant presence, analyzed using binary and continuous treatment variables. In both cases, we observe an average of

around 2 percent wage increase in Roraima post-treatment when a Venezuelan was working in similar occupations or industries. Columns (3) and (4) show results for industries and occupations without any Venezuelan workers recorded in our data. Here, the wage effects were slightly higher for economic activities using the binary treatment version and lower (though less precise) in the continuous treatment version. For occupations, the results showed almost a 3 percent average wage increase.

These findings suggest that the positive effects observed in the aggregate market are mirrored in our economic sector analysis. Specifically, sectors and occupations with a high concentration of immigrants saw a marginally lower average wage increase compared to those without immigrants.

In summary, the entry of immigrants into Roraima appears to have generally driven up wages, likely due to increased labor demand and increased presence of Venezuelan workers in informality. However, those immigrants who integrated into the formal labor market exerted a slight downward pressure on native wages. This effect was not solely based on education levels but also the types of occupations and economic sectors, as evidenced by the immigrants' sorting patterns in the labor market.

Table 2.5: Heterogeneous Treatment Effects by Activity and Occupation

	Any Immigrants		No Immigrants	
	(1)	(2)	(3)	(4)
Economic Activity				
Treat: Binary	0.019***		0.023***	
	(0.005)		(0.007)	
Treat: VZ Ratio x 100		0.019***		0.014**
		(0.002)		(0.005)
Occupation				
Treat: Binary	0.022***		0.027***	
	(0.005)		(0.009)	
Treat: VZ Ratio x 100		0.021***		0.030***
		(0.003)		(0.005)
Individual FE	X	X	X	X
Year FE	X	X	X	X
R2 Adj.	0.855	0.855	0.931	0.931
RMSE	0.241	0.241	0.200	0.200
N Clusters	53	53	53	53
N	534 212	534 212	34 093	34 093

[1] Standard-errors are clustered by municipality.

[2] Propensity score explanatory variables are gender and race indicators, age, age-squared, tenure, tenure-squared and education level.

[3] * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Substitution through Occupation Changes of Brazilian Workers

To establish more robustness to our assumption of small but significant substitution effects happening in the formal sector, we can examine the job changes of native workers in response to immigrant sorting. This is similar to the approach taken in Foged and Peri, 2016, which assumes that refugees entering the labor market took on manual-intensive jobs, potentially allowing native workers to shift into other occupations.

To analyze the effects of immigration on native workers' job choices, we create a binary variable to indicate whether an individual in occupations where immigrants are observed ever changed their occupation to a position where we do not observe Venezuelans. Similar to previous frameworks, we compare these occupational changes across the control and treatment groups.

We use this variable as the dependent variable in Equation (2.1) and (2.2), presenting the regression results in Table 2.6. They correspond to a linear probability measurement of an individual being a "mover" from immigrant occupations to non-immigrant ones.

Table 2.6: Movement from Occupations with Any Immigrants to Occupations with No Immigrants

	Movement (1)	Movement (2)
Treat: Binary	0.002***	
	(0.000)	
Treat: VZ Ratio x 100		0.001***
		(0.000)
Individual FE	X	X
Year FE	X	X
R2 Adj.	0.008	0.008
RMSE	0.072	0.072
N Clusters	53.000	53.000
N	576 800	576 800

[1] Standard-errors are clustered by municipality.
[2] Propensity score explanatory variables are gender and race indicators, age, age-squared, tenure, tenure-squared and education level.
[3] $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

The results indicate that, overall, there was a small but significant change of occupation for Brazilians conditional on the presence of Venezuelans. This implies that even though the substitution effect is measurable, it should be considered too small to provide any significant overall effect in the market.

2.6.2 Relationship between Formal Sector Effects and Informal Sector Effects

So far, our analysis has demonstrated the positive labor market effects of the Venezuelan immigration crisis. We also showed that individuals not involved in occupations or economic sectors and occupations with a sizeable Venezuelan presence experienced more prominent wage increases compared to their counterparts. Moreover, we did not observe any negative wage effects for Brazilians in manual labor occupations where immigrants were mostly concentrated.

One way to interpret these results is to consider the role of Venezuelan immigrants who are not part of the formal market. We showed that many Venezuelan immigrants in the region had sought refugee status, but only a small percentage entered the formal market. This potentially implies that many of them worked informally or were in seek of employment. Informal workers tend to have lower levels of education and they specialize in manual tasks, while those in the formal labor market often hold cognitive or technical jobs, with some overlap.

In Roraima, about 45 percent of the workforce is involved in the informal labor market. To analyze whether there are negative impacts on wages in the informal labor market, we use the PNAD-C dataset from 2012-19. PNAD-C is a representative household survey conducted by the Brazilian Institute of Geography and Statistics (IBGE) every quarter that includes a variety of socioeconomic information, such as employment status, income, race, gender, and education level among others. However, a major limitation of the dataset is that, unlike RAIS, social identification of firms and persons, and nationality, are not observable. Therefore, we cannot directly analyze the impacts on Brazilian citizens or use individual fixed effects to control for time-invariant individual characteristics. Moreover, unlike RAIS, we cannot observe the municipality of individuals. As a result, estimates from PNAD-C are potentially downward biased. The adapted model is specified as follows:

$$y_{ist} = \beta D_{st} + f(X_{it}) + \theta_s + \alpha_t + \varepsilon_{ist} \tag{2.7}$$

where y_{ist} represents the logarithmic average wage that individual i earned in quarter t in state s. D_{ist} is the indicator function that becomes one if individual i is from the state of Roraima after 2013. $f(X_{it})$ is the covariate matrix linear function, including education, race, gender, age, and age-squared. State fixed effects and year fixed effects are represented by θ_s and α_t, respectively. The term β estimates the effects of the Venezuelan refugee crisis in Roraima on logged wages. The error term ε_{ist} is clustered at the state level, given we cannot observe municipality in the PNAD-C data.

Table 2.7: Log Wage Effects in the Informal Sector

	Informal Log Wages		
	(1)	(2)	(3)
Treat	−0.018	−0.022	0.012
	(0.029)	(0.027)	(0.029)
Immigrant Occupations		0.077***	
		(0.016)	
Treat × Immigrant Occupations		0.011	
		(0.019)	
Immigrant Activities			0.140***
			(0.012)
Treat × Immigrant Activities			−0.085***
			(0.017)
N	67 894	67 894	67 894
Year FE	X	X	X
State FE	X	X	X

[1] Standard-errors are clustered by state.

[2] Propensity score explanatory variables and covariates in the main spacification are gender and race indicators, age, age-squared, tenure, tenure-squared and education level.

[3] * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2.7 presents the outcomes of our analysis using equation (2.7). Column (1) shows the aggregate treatment effect result, at around -1.8 percent, which is not statistically significant. Despite the inherent coarseness of the PNAD-C survey data, it still provides limited evidence of a substitution effect in the informal market due to the increased labor supply in Roraima.

Moreover, PNAD-C includes data on economic activities and worker occupations, albeit at a lower level of detail. To assess the impact of immigrant exposure on specific industries or job occupations more accurately, we link sectors with a higher immigrant presence in both variables in RAIS to corresponding categories in PNAD-C. We assume that Venezuelan immigrants have similar preferences regarding economic activities and occupations in both informal and formal sectors. The categories used for the indicator function are general service, crafting, and construction helpers for job occupations, and construction, restaurants, and commerce for economic activities.

The model, similar to Equation (2.7), differs by allowing the heterogeneity dummy to interact with the treatment indicator. Columns (2) and (3) in Table 2.7 show results for the informal sector; specifically, the interactions with occupations and economic activities, respectively.

In the informal sector, the data indicate that both occupations and firm activities were associated with higher wages compared to other categories. However, with the treatment interaction, the results become statistically insignificant for occupations and show a significant 8.5 percent decrease for firm activities. This suggests that immigrants may have adversely affected the earnings of individuals in these categories in Roraima post-crisis.

The results indicate a significant impact of immigration on the informal sector, highlighting a substitution effect that adversely influences wages. Conversely, in the formal sector, the analysis suggests a potential benefit for workers, attributable to the complementarity effect that the informal sector provides to formal employment and a possible increase in efficiency within the formal market.

2.7 Conclusion

In this paper, we conducted a comprehensive analysis of the labor market impacts of the Venezuelan crisis in Brazil, focusing on the state of Roraima, where the crisis had a direct impact. The state's geographical isolation helps us use the immigration shock as a natural experiment, allowing us to measure its effects on the local labor market. Using a difference-in-differences model, we explored the potential differences in effects based on market diversity in terms of economic sector and worker occupation.

Our findings reveal that the average monthly wages in Roraima increased by approximately 2 percent in the early stages of the crisis. By analyzing the presence of Venezuelan formal workers by firm's economic sectors and occupation type, we found that Brazilian workers involved in economic sectors and occupations that had a presence of Venezuelan immigrants experienced a slightly lower wage increase than those involved in sectors and occupations without any Venezuelan presence. We also observed no significant changes in formal employment displacement of Brazilians but did find evidence of them moving from high-immigrant occupations to low-immigrant occupations in the post-treatment years.

Furthermore, using nationally representative survey data, our analysis of the informal labor market revealed that while Brazilian informal workers did not experience a significant drop in wages on aggregate, those involved in sec-

tors with the presence of Venezuelan immigrants experienced a significant wage drop. We can conclude that immigrants in the informal labor market acted as complements to the formal labor market, allowing the overall formal wage to increase and offsetting any substitution effect of foreign workers within formality.

In summary, our study emphasized the need to consider the various factors that can influence the impacts of refugees on the labor market. While our research suggested that refugees can bring benefits, it also highlighted the potential drawbacks of large-scale immigration in regions with a significant informal economy. Moving forward, policymakers should prioritize policies that improve the welfare of refugees and promote their active participation in the economy while also being mindful of the potential negative effects on those working informally. Future research should further explore the impacts of the population boom in Roraima, including improving the understanding of the effects on consumption, health, and public safety outcomes.

Chapter 3

Assortative Matching and the Gender Wage Gap

3.1 Introduction

Human capital has been a crucial component in narrowing wage disparities between men and women. In the United States and similar developed and emerging markets, women are more likely than men to hold college degrees (OECD, 2024). However, there is still substantial gap, and several studies attempted to tackle the issue under different perspectives. Differences in productivity has been argued by Mulligan and Rubinstein (2008). There is also evidence that non-monetary factors, such as preferences for flexible hours, play a major role in generating the gap (Goldin, 2014).

A recent strand of the literature focus primarily on the contribution of firm-specific pay policies that would be important in creating differences across genders. Building on a simple rent-sharing model, these papers break down the wage into two main components: a worker component, solely generated by human capital levels and other worker characteristics, and a firm component, arising from firm heterogeneity such as economic activity, market power, and size. Leveraging from administrative data, Card, Cardoso, and Kline (2016) (CCK) introduced a Kitagawa-Oaxaca-Blinder (KOB) decomposition (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) to measure the contribution of firm effects on the gender wage gap, finding that around a fifth of the gap arises from firm premiums. Even though analyses based on rent-sharing models are effective in providing a comprehensive overview of the impact of firm-specific pay premiums on the gender wage gap, these models typically assume that the value of worker characteristics remains constant across firms and vice-versa. Therefore, this "additive separability" assumption is restrictive as it constrains the ability to

capture particular worker-firm interactions due to the rank condition. This precludes scenarios where comparative advantages[5] arise from specific worker-firm matches. These models may fail to account for an important source of wage variation when different classes of firms perceive similar workers differently, or when particular matches in the labor market are advantageous to certain workers, which could significantly contribute to explaining gender wage disparities.

[5] In this paper I use "comparative advantage" effects and "complementarity effects" interchangeably.

In this paper, I provide the first comprehensive analysis of worker-firm interaction effects on the gender wage gap, explicitly accounting for assortative matching in the labor market. Extending the two-step distributional framework of Bonhomme, Lamadon, and Manresa (2019), I apply k-means clustering and a Gaussian mixture model to the log-hourly wage distribution from massive linked employer-employee data in Brazil. The data provides the universe of formal workers and firms, with a rich set of variables, such as extremely detailed economic activity and occupation codes, gender, race, education level, firm location and more.

My innovative approach groups workers and firms into "types" and "classes" respectively, reducing dimensionality to satisfy the rank condition necessary to explore worker-firm match effects in the wage structure. The model assumes that each group represents categories of workers and firms that are comparable and, when interacted, generate wages by drawing from a Gaussian distribution where parameters are specific to that match. This methodology allows for a wage generation process that deviate from the restrictive additive separable framework, enabling the identification of wage effects that arise solely from specific worker-firm interactions.

For firms, I employ k-means clustering to group them under similar payment distributions. To determine the optimal number of clusters, I utilize a gap statistics analysis (Tibshirani, Walther, and Hastie, 2001), which identifies the point at which within-cluster variance is minimized. For workers, I model the probability density function (PDF) of wages within each firm class as a mixture of log-normal distributions. I demonstrate the robustness of my results across alternative specifications of these combinations. Furthermore, I provide evidence that estimated clusters can be mapped to observable characteristics, validating their economic significance. Since the identification of my model relies on job movers, I test the exogenous mobility assumption by showing job movement is not related to unobservables.

My model is flexible enough to allow for the identification of differential firm valuations of workers with similar unobserved characteristics. Through Monte Carlo simulations, I identify three distinct channels that contribute to gender wage disparities.

My key contribution is the identification of a "match effect". This component captures the wage effect of specific worker-firm interactions, revealing complementarities that arise when particular worker types are matched with certain firm classes. I simulate a labor market with no complementarity, therefore absent of match effects, and compare it with baseline estimates. While typical separable models struggle to identify this component, I find that women are less likely to benefit from positive complementarity effects in wages. Quantitatively, in a counterfactual world without complementarities, women's average log hourly wages increase by one log point (from 2.10 to 2.11), while men's decrease by three log points (from 2.33 to 2.30). Hence, transitioning from the observed labor market with complementarities to a simulated market without reduces the gender wage gap in log hourly wages from 0.24 to 0.20, a decrease of approximately 17 percent. This result suggests that female workers are more likely to be found in disadvantageous worker-firm interactions that yield negative complementarities in wages and even when they are present in interactions that yield positive complementarities, women tend to benefit less than men.

My results also indicate that the complementarity effect grows with higher levels of human capital and the complexity of occupations. These contributions increase with both education and age. Notably, workers in occupations typically associated with the hospitality industry, such as cleaners and waiters, show no evidence of complementarity effects. However, for individuals in occupations requiring STEM degrees, such as engineers and economists, complementarity contribution becomes positive. For managers, it accounts for as much as one-third of the gender wage gap.

In the spirit of Card, Cardoso, and Kline (2016), I also explore the overall contribution of firms to the gender wage gap by assuming the labor market is under assortative matching[6] and firms evaluate worker characteristics on top of offering premiums. To perform this analysis, I hold constant the distribution of worker clusters across men and women, measuring the counterfactual wage difference when the distributions of firms and their expected payments vary. This approach reveals two components, in addition to the match component, that mirror the established literature: sorting (women's under-representation at higher-paying firms) and bargaining (equally productive women receiving a smaller share of payments). Additive separable models potentially underestimate these effects since they assume the value of worker characteristics is constant across the labor market, thereby imposing a downward bias to the impact of assortative matching in generating the gender wage gap.

The sorting component, representing the contribution of differences in firm allocations in the labor market, accounts for approximately 37.5 percent of

[6] For this paper, I follow (Becker, 1973) to consider assortative matching as the propensity of high quality firms to match with high quality workers.

the 24 log point gender wage gap in Brazil. This sorting effect is substantially larger than estimates obtained from additive separable models, about 9 percent. The enhanced magnitude stems from my model's ability to capture heterogeneous firm-specific returns to worker characteristics. Specifically, it reveals that women are disproportionately concentrated in labor market segments where firms offer lower returns to worker and firm characteristics for all workers, regardless of gender. This component is less relevant for young individuals, but increases considerably for older and highly educated individuals, reaching about 40 percent to college graduates.

The final component, representing the contribution of differential payments to similarly productive men and women, accounts for approximately 8.3 percent of the gender wage gap. This "bargaining" effect suggests that even when women overcome sorting barriers, they still face wage disadvantages within firms. Collectively, the "match," "sorting," and "bargaining" components explain more than sixty percent of the observed gap, suggesting that understanding assortative matching in the labor market is essential to mitigate wage disparities.

This paper belongs to the applied literature investigating the channels generating the gender wage gap. While the gap narrowed in recent decades (Blau and Kahn, 2008) due to increases in female human capital (Goldin, Katz, and Kuziemko, 2006; Black, et al., 2008; Ceci, et al., 2014), Goldin (2014) finds that women's preference for flexible hours over monetary compensation is a relevant factor to narrow the remaining gap. Bertrand, Black, et al. (2019) shows that women are disproportionately underrepresented in jobs with high returns on human capital investment.

The more recent strand in the literature investigates the contribution of firms to the gender wage gap. These papers belong to the applied literature that employs the AKM model (Abowd, Kramarz, and Margolis, 1999; Card, Schmutte, and Vilhuber, 2023), specifically focusing on firm effects on the gap. Generally, these studies use linked employer-employee data in a two-sided separable model with worker and firm identifiers as "plugin" estimators in a log wage linear regression. Card, Heining, and Kline (2013) showed wage dispersion could be largely attributed to these firm components using West Germany data and Card, Cardoso, and Kline (2016) proposed a KOB decomposition using Portuguese data, finding that firms contribute around 20 percent to the gender wage gap [7].

Nevertheless, the AKM-KOB analysis assumes human capital returns for specific workers are constant across the labor market. It also requires special data manipulation to extract a dual connected set of firms through male and female

[7] Other papers using the AKM model to explore firm effects KOB decomposition on the gender wage gap: Gallen, Lesner, and Vejlin (2019) in Denmark, Bruns (2019) in West Germany, Jewell, Razzu, and Singleton (2020) in the UK, Masso, Meriküll, and Vahter (2022) in Estonia, and more recently Casarico and Lattanzio (2024) in Italy.

workers changing jobs (Abowd, Creecy, and Kramarz, 2002; Card, Cardoso, and Kline, 2016). I show in my supplementary material that this data restriction may not be innocuous, since the trimming procedure may disproportionately preserve larger firms, which typically exhibit higher wage dispersion. Moreover, wage variance analysis in AKM can be biased by underestimating the role of worker-firm interactions[8]. It was first assumed to be an economic phenomenon; however, Andrews, et al. (2008) showed that this was, in fact, an econometric issue related to small sample bias. The proposed straightforward correction to this "limited mobility bias" can be computationally prohibitive (Gaure, 2014; Azkarate-Askasua and Zerecero, 2023). Kline, Saggio, and Sølvsten (2020) introduced a Leave-One-Out methodology to fix it.

[8] Some earlier examples of biased effects are Barth and Dale-Olsen (2003) using Norse data and Gruetter and Lalive (2009) using French data, where they found negative covariance estimates in joint worker-firm effects.

My paper goes in line with alternative approaches that moves away from AKM's additive separable assumption to avoid biased results and capture match effects. Woodcock (2008) proposed a random effects approach to satisfy the rank condition. More recently, Bonhomme, Lamadon, and Manresa (2019) (BLM) proposed a novel approach that involves clustering both firms and workers into broader categories. This method offers two key advantages. First, reducing the worker firm dimensions is computationally tractable and allows for further exploration of worker-firm match effects. Second, its non-separable nature gives a unique opportunity to observe how firms valuate workers differently under similar circumstances but differing only by gender. Bonhomme, Holzheu, et al. (2023) demonstrated that random effect models such as the BLM are particularly effective in circumventing the AKM limitations, even in short time panels.

I contribute to the literature in many ways. First, to my knowledge, this study is the first to implement the methodology of Bonhomme, Lamadon, and Manresa (2019) in the context of analyzing wage disparities between two groups. Secondly, I empirically show that separable models underestimate the role of firms in generating the gap. More importantly, I find that some interactions in the labor market exhibit comparative advantage effects, generating wage levels that substantially exceed predictions from the traditional models, but with less intensity when these interactions occur with female workers. These matches contribute significantly to wage disparities but are often smoothed out under the separability assumption. Additionally, I find these interactions exist particularly in high-paying, larger firms and they are particularly strong in highly educated workers[9].

[9] Studies that explore high paying firms and top earners are, for example, Bertrand, Goldin, and Katz (2010) and Bertrand, Black, et al. (2019), demonstrating the existence of a "glass ceiling" effect.

My findings have meaningful policy implications. They show that closing the gender wage gap requires improving pay practices in key roles where highly skilled women are employed, particularly in leadership positions. Equally im-

portant are efforts to break down barriers in the labor market that push women into lower-paying firms. These firms not only offer smaller wage premiums but also limit the returns on women's skills and education, deepening income inequality over time.

The remainder of the paper is organized as follows: Section 3.2 provides an explanation of what is a complementarity effect and why additive separable models cannot capture it under typical settings. Section 3.3 provides an overview of the Brazilian data used in this study. Section 3.4 explains the BLM clustering method in two-steps. In Section 3.5, I provide the clustering results. In Section 3.6, I construct the Monte Carlo simulation counterfactuals. I conclude the paper in Section 3.7.

3.2 Additive Separable Models and Complementarity

Researchers are often interested in identifying the returns to unobserved heterogeneity of both workers and firms in labor markets, particularly when administrative data with social identifiers are available. In many empirical studies, these social identifiers are utilized as "plug-in estimators," commonly modeled as fixed effects in a linear equation where the outcome is the wage in logs.

For instance, consider a labor market consisting of N workers and J firms, where workers and firms interact over T periods. Under the assumption of additive separability, the log wage w of worker i at time t, net time varying effects, can be modeled as:

$$\log w_{it} = \alpha_i + \phi_j + \varepsilon_{it} \tag{3.1}$$

$$\text{s.t. } j = J(i,t) \tag{3.2}$$

where α_i represents the fixed effect associated with worker i (capturing unobserved worker-specific characteristics such as skills or human capital), and ϕ_j denotes the firm-specific premium associated with firm j (the wage component determined by firm characteristics, independent of worker-specific attributes). The error term ε_{it} captures idiosyncratic shocks. Firm assignment is indicated by the function $J(i,t)$, which tracks the firm employing worker i at time t.

In this framework, the additive separable model assumes constant returns for both workers and firms. That is, the firm-specific premium ϕ_j is unaffected by the characteristics of the worker employed by the firm, and vice versa. This implies that reshuffling workers across firms does not alter the firm component

of wages. Such an assumption is particularly strong and potentially unrealistic in labor markets where comparative advantage is thought to play a role in worker-firm interactions (Shimer and Smith, 2000; Eeckhout and Kircher, 2018).

To allow for complementarity between workers and firms, one could extend the model to include interaction effects. Specifically, the wage equation can be rewritten as:

$$\log w_{it} = \alpha_i + \phi_j + M_{ij} + \varepsilon_{it} \tag{3.3}$$

$$\text{s.t. } j = J(i,t) \tag{3.4}$$

where M_{ij} represents the interaction effect between worker i and firm j. This term captures the potential complementarity effect that only arises when worker i is employed at firm j. It reflects the idea that certain worker-firm pairings generate higher (or lower) wages than what would be predicted based solely on the worker's fixed effect α_i and the firm's premium ϕ_j. However, estimating this interaction term in practice is infeasible due to the large dimensionality of the model. The matrix M_{ij} would have $N \times J$ terms, which quickly becomes computationally intractable given that linked employer-employee data typically contain millions of workers and thousands of firms.

Moreover, these models are prone to bias in settings with short panel datasets, where estimating the worker and firm fixed effects becomes difficult. Building on the approach of Bonhomme, Lamadon, and Manresa, 2019, I employ a methodology that reduces the dimensionality of workers and firms by clustering them into latent groups. This allows for the estimation of complementarity effects while avoiding the rank deficiency problem inherent in models with such high-dimensional interactions. Specifically, workers and firms are grouped based on their interactions in the labor market. I assume each worker-firm interaction draws wages from log-normal distributions, meaning I can relax the linear assumption and the additive separability.

The wage generation function now can be expressed as:

$$\log w_i = f(\alpha_{L(i)} \mid \phi_{K(i,j)}) \tag{3.5}$$

where $L(i)$ denotes the assignment function of worker i's type, and $K(i,j)$ denotes the assignment function of firm j's class. $f(\alpha_{L(i)} \mid \phi_{K(i,j)})$ denotes a probabilistic function that draws wages from a log-normal distribution that is specific related to the match of worker type l and firm class k. Since wages are assumed to be derived specifically from worker-firm interactions, some matches

are allowed to yield comparative (dis)advantage effects on wages, given that returns to firm and worker characteristics, under this model, are not necessarily constant across the labor market.

Therefore, recovering these latent types can be done by exploring the surface of observed wages. The idea behind the model is to recover Gaussian distributions that are "combined" in the full distribution, derived from all the matches occurring from the labor market. Thus the Gaussian mixture model application.

A key assumption for this model is the exogenous mobility assumption, meaning job mobility depend on the type of the worker and the classes of the firms, but not directly on earnings. I discuss the exogenous mobility assumption in my model in Section **??**.

In the context of wage differences due to gender, my approach is to estimate the Gaussian mixture parameters in a pooled dataset, meaning the model does not observe gender at first. The reasoning is to facilitate the comparability of individuals under the same umbrella of unobserved heterogeneity. I discuss further the model in Section 3.4.

3.3 Data

In this section, I provide an overview of the Brazilian administrative data used for the study and the preparatory cleaning for the analysis, followed by a descriptive statistics of the cleaned sample.

3.3.1 Data Overview and Institutional Background

I use the *Relação Anual de Informações Sociais* (RAIS), an extensive linked employer-employee dataset (LEED) from Brazil spanning from 2010 to 2017. RAIS is mandated and maintained by the Brazilian Ministry of Labor and Employment, serving as a source for the administration of tax and social programs. The dataset offers an universal representation of the formal labor market in Brazil and is characterized by its richness in variables.

A key advantage of using the *RAIS* dataset for this analysis is the relative homogeneity of job-related amenities across firms due to Brazil's robust labor regulations. The Brazilian labor laws, known as the Labor Laws Consolidation (*Consolidação das Leis do Trabalho* (CLT), in Portuguese), mandates a broad range of standardized benefits and protections for all formal workers, regardless of industry or firm size. This regulatory framework significantly reduces variation in non-wage compensation, allowing the analysis to focus more cleanly

on wage differentials without the confounding effects of divergent job-related amenities.

For example, Brazilian law requires all formal employees to receive the 13th salary, which is essentially a mandatory annual bonus equivalent to one month's wage, usually paid during Christmas time. Additionally, firms are obligated to provide meal vouchers or food stipends, as well as transportation subsidies for commuting. These benefits are non-negotiable and standardized across the formal labor market. Moreover, formal workers are entitled to thirty days of paid vacation, overtime pay, and severance protections via the *Fundo de Garantia por Tempo de Serviço*[10] (FGTS), which further ensures that variations in non-wage job characteristics can be minimized.

[10] Roughly translated as Severance Indemnity Fund for Length of Service

In Brazil, maternity leave is a legally guaranteed right under the CLT. Female employees are entitled to 120 days of paid maternity leave, funded by the Brazilian Social Security system. In some cases, companies can extend this leave to 180 days through the Empresa Cidadã program, which offers tax incentives to employers. During maternity leave, the employee's job is protected, and she is guaranteed to return to her position or a similar one without loss of salary or benefits. Additionally, Brazilian law prohibits the dismissal of pregnant workers from the moment pregnancy is confirmed until five months after childbirth, with some exceptions under fair cause.

This regulatory uniformity is particularly beneficial for my analysis, as it mitigates concerns that differences in firm payment patterns are due to job amenities that could ultimately explain wage differentials between male and female workers. In contrast, in countries where non-wage compensation varies significantly across firms or sectors, disentangling wage differences from benefit-driven compensation can complicate the analysis of wage gaps.

In my study I focus on São Paulo state, which represents the most economically dynamic region in Brazil, making sure my results are not driven by geographical heterogeneity. For example, a male worker in manufacturing and a female worker in retail, though in distinct sectors, would both receive a standardized package of legal protections and benefits coming not only from federal law, but also from local state law, ensuring that wage comparisons are not distorted by differences in state policies.

Regarding gender dynamics in São Paulo's labor market, it is important to note that, similar to other countries analyzed in the literature, approximately more than 50 percent of the Brazilian women there participate in the labor force, with 71 percent of these women employed full-time. This proportion rises to 90 percent when considering only those employed in the private sector. Furthermore, the gender wage gap in Brazil mirrors those observed in more devel-

oped economies, offering additional comparative insights. In 2016, the median earnings gap between male and female full-time workers was approximately 14.3 percent in Brazil, closely aligned with the average of 13.4 percent observed across OECD countries, and slightly better than the 18.1 percent reported for the United States (OECD, 2024).

3.3.2 Data Preparation

The RAIS database records each formal employment contract as a separate entry, meaning that for any given year, a worker with multiple contracts, whether with the same employer or different firms, will appear multiple times. To address this, and following the methodologies of Gerard, et al. (2021) and Lavetti and Schmutte (2023), I refine the dataset by retaining only the longest-duration and highest-paid contract for each individual per year. This adjustment shifts the data from a contract-year structure to an individual-year framework, ensuring that the analysis focuses on each worker's primary employment.

To align with a long-run perspective, the sample is further restricted to a *quasi*-full-time workers, defined as those working a minimum of 30 hours per week, and limited exclusively to the private sector. I allow this flexibility to capture a certain degree of non-monetary preference particularly found in female cohorts (Goldin, 2014). This exclusion criteria eliminates part-time employees, public sector workers, and the self-employed from the analysis, thereby focusing on a more homogeneous labor market.

Biennial Grouping and Panel Balancing

The organization of the data for my analysis involves grouping the dataset into jumping biennials. Specifically, the years 2010 and 2012 are paired, 2011 and 2013, and so forth. This method skips intermediate years to avoid transitional anomalies that may occur in short periods, such as firm mergers or changes in identifiers. This "jumping" approach closely mirrors the sample selection method employed by Bonhomme, Lamadon, and Manresa (2019).

In my analysis, it comprises of six sets of balanced panel data spanning from 2010 and 2012 to 2015 and 2017. Each biennial set is balanced and analyzed to estimate worker and firm clusters, with final estimates related to wages presented as a weighted average of these samples. This "rolling" approach has been used to some extent in Card, Cardoso, and Kline (2016) and Lachowska, et al. (2023).

Each biennial panel is balanced, ensuring that the same set of workers and firms are observed consistently within each two-periods. In addition, firms with pronounced gender preferences are excluded from the analysis. Only firms

exhibiting a gender ratio of 1 to 4 are included, which helps mitigating any potential bias that could arise from firm gender imbalance.

3.3.3 Summary Statistics

Table 3.1 reports descriptive statistics by gender cohorts for the aggregated cleaned data, representing the first year of each biennial sample. Columns (1) and (2) represent the statistics for female and male workers, respectively.

The dataset encompasses a total of 346,617 unique firms. Of these, a substantial portion is relatively large; approximately 204,994 firms employ 10 or more workers, and 58,866 firms have at least 50 workers. The average firm size across the sample is 57 employees, but the median firm size is considerably smaller, at 13 employees, indicating a skewed distribution.

Gender related educational attainment confirms that women are generally more educated than their male counterparts. The data show a higher prevalence of men without high school diplomas, while women are more likely to have completed high school or pursued some college education. As stated previously, this educational dynamic is consistent with recent trends observed in both developed and developing nations, such as the United States and other OECD countries.

Approximately 40 percent of the female sample is under 30 years old, with another 50 percent aged between 31 and 50. In contrast, 37 percent of the male sample is under 30, with 49 percent in the 31 to 50 age bracket. Moreover, men are slightly more represented in the over-50 cohort, constituting 12 percent compared to 8 percent of women. Hence, the average experience in the labor market is 4.6 years for males and 4.0 years for females.

Industry distribution varies significantly between genders. Men dominate in sectors such as manufacturing, agriculture, and trade, whereas women are predominantly engaged in services, an umbrella term that includes sectors such as healthcare, education, hospitality, and financial services.

The occupational distribution also highlights a notable gender sorting: women are almost twice as likely as men to hold administrative positions, representing 34 percent of women compared to 18 almost of men. Men are more frequently employed in manual labor-intensive roles such as in agricultural settings and factories.

Despite these occupational disparities and the educational advantages observed for women, the unweighted gender wage gap remains substantial at approximately 23 log-points. This gap persists even though women are, for instance, equally likely as men to occupy scientific roles, which typically require higher educational qualifications.

Note: [1]Descriptive statistics calculated from the first year of each biennial sample (2010-2015). [2] Percentages may not sum to 100% due to rounding. [3]The number of firms is the same for both genders since every firm in the cleaned sample employs both male and female workers.

Features	**Female Workers** (1)	**Male Workers** (2)
Firm Characteristics		
Number of Firms	346 617	346 617
Firms with $\geq$ 10 Workers	204 994	204 994
Firms with $\geq$ 50 Workers	58 866	58 866
Mean Firm Size	57	57
Median Firm Size	13	13
Worker Characteristics		
Education (%)		
Dropout	22	28
High School Graduates	48	45
Some College	30	27
Age (%)		
$<$ 30	40	37
31–50	50	49
$\geq$ 51	10	14
Sector of Employment (%)		
Primary	2	2
Manufacturing	19	26
Construction	1	2
Trade	24	25
Services	54	45
Occupation (%)		
Scientific and Liberal Arts	11	11
Technicians	11	11
Administrative	34	18
Managers	5	7
Traders	25	22
Rural	1	2
Factory	13	29
Labor Market Outcomes		
Mean Tenure (years)	4.04	4.63
Mean Log-Wage	2.06	2.29
Variance of Log-Wage	0.52	0.65
Worker-Year Observations	9 503 233	10 283 471
Unique Number of Workers	3 497 651	3 725 990
Gender Fraction (%)	48	52

3.3.4 Extended Mincer Equation

As a first step to analyze the gender wage gap, I provide a classical Kitagawa-Oaxaca-Blinder (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973) decomposition of an extended Mincer equation and an AKM equation, assuming the gap is a mean difference of female and male wages. A "Mincer wage function" can be specified as:

$$w_{it} = \beta_0 + \beta_1 \text{Age}_{it} + \beta_2 \text{Age}_{it}^2 + \beta_3 \text{Education}_{it} + \beta_4 \text{Occupation}_{it} + \beta_5 \text{Activity}_{it} + \varepsilon_{it} \tag{3.6}$$

where w_{it} is the natural logarithm of hourly wages for individual i in time period t, regressed on the worker's age and their squared age, their education level, the firm's industry, the worker's occupation, and a idiosyncratic error term. For the Oaxaca decomposition, I run this regression for the male and female observations separately, for each biennial sample.

Assume the matrix of explanatory observables can be expressed as X^g, where g represents the gender sample used in the regression. Also assume β is the vector of estimates. The KOB decomposition can be expressed as:

$$\bar{w}^m - \bar{w}^f = \underbrace{\left(\bar{X}^m - \bar{X}^f\right)\hat{\beta}^f}_{\text{Explained}} + \underbrace{\bar{X}^f\left(\hat{\beta}^m - \hat{\beta}^f\right)}_{\text{Unexplained}} \tag{3.7}$$

where $\left(\bar{X}_m - \bar{X}_f\right)\hat{\beta}_f$ represents the "explained" component of the decomposition. In simpler terms, this term represents a counterfactual scenario where men and women possess the same returns to covariates, however, they differ in these covariates' distribution. The unexplained component, on the other hand, captures differences in the returns to these characteristics. This is expressed as $\bar{X}_f\left(\hat{\beta}_m - \hat{\beta}_f\right)$, where the difference in coefficients $(\hat{\beta}_m - \hat{\beta}_f)$ measures a scenario where men and women have the same observable characteristics, however, the market values differently each gender. The unexplained portion is often interpreted as the part of the wage gap that cannot be accounted for by observable factors alone, potentially indicating discrimination or other structural labor market imbalances.

Table 3.2 presents the overall log hourly wage gap in means, the explained, and the unexplained portion of the gender wage gap across the six biennial samples, along with the number of observations for each sample. The overall wage gap remains consistent at 24 log-points for the first three samples. However, the gap slightly decreases in the subsequent samples, with the smallest gap observed in 2015-2017 at 22 log-points.

Table 3.2: Extended Mincer Equation KOB Decomposition For Each Biennial Sample

Sample	Overall Gap	Explained Gap	Unexplained Gap	N
2010–2012	-0.244	-0.0651	-0.179	5 946 240
2011–2013	-0.244	-0.0637	-0.180	6 145 676
2012–2014	-0.244	-0.0642	-0.180	6 534 444
2013–2015	-0.241	-0.0621	-0.179	6 787 446
2014–2016	-0.230	-0.0571	-0.173	7 086 062
2015–2017	-0.221	-0.0542	-0.167	7 073 540
Weighted Avg[a]	**-0.237**	**-0.0611**	**-0.176**	39 573 408[b]

Note: [a]Weighted average calculated using sample sizes as weights and the gap as $female - male$. [b]Total number of observations across all samples. [1]Extended Mincer equation defined as $\log(y_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \beta_3 \text{Education}_i + \beta_4 \text{Occupation}_i + \beta_5 \text{Activity}_i + \varepsilon_i$. [2]Explained gap represents differences in distribution of characteristics. Unexplained gap represents differences in estimated returns to characteristics.

In this setting, the explained portion of the Oaxaca decomposition accounts for approximately 6.11 log-points, or roughly one-quarter of the total gender wage gap. This indicates that observable factors, such as the allocation of workers across different occupations or sectors, explain about 25 percent of the wage differential in an additively separable labor market.

In Section 3.6, I extend the analysis by introducing firm identifiers as fixed effects under an AKM framework following Card, Cardoso, and Kline (2016). Under this specification, firm effects explain about 9 percent of the gender wage gap.

The issue with separable models is the assumption that these components should not vary depending on the association happening. Under AKM, these firm effects will occur in any worker reshuffling instance of the labor market.

In the next sections, I propose the distributional framework of Bonhomme, Lamadon, and Manresa (2019) to capture particular interactions in the labor market that does not necessarily follow an additive separable assumption.

3.4 Empirical Framework: The BLM Model

Estimating the Gaussian mixture requires two main parts. Following Bonhomme, Lamadon, and Manresa (2019), I assume cluster membership of firms is exogenous to the model, allowing their estimation by employing straightforward clustering methods from features observed from the data. Still following

BLM, I choose to cluster firms based on their wage cumulative distribution function using k-means clustering (MacQueen, 1967).

In the second part, I take the estimated firm clusters, called "firm classes", to assume that they are Gaussian mixtures of latent worker types in log wages. In the spirit of AKM settings, I leverage individuals changing jobs to identify the Gaussian parameters.

Finally, I use a *maximum a posteriori* estimation to find the most likely worker type for each worker observation. After the classification, I split the data into male and female cohorts.

3.4.1 Recovering Firm Classes

The first objective is to recover firm clusters, or "firm classes", which are initially unobserved in the data. The approach relies on two key assumptions. First, the mapping of firms to clusters is exogenous to labor market dynamics.

Formally, let $k(j)$ denote cluster assignment of firm j. The exogeneity assumption can be expressed as:

$$P(k(j)|X) = P(k(j)) \tag{3.8}$$

where X represents labor market conditions and worker characteristics. In plain language, this condition ensures that the probability of a firm belonging to a firm class is unconditional to these labor market features, which allows a direct estimation of firm classes using the clustering method.

Secondly, the wage distribution in the data follows a log-normal shape for workers, conditional on these firm clusters. Consequently, each firm class represents a Gaussian mixture of log-wages. Within these mixtures, each component corresponds to a log-normal distribution arising from the unobserved heterogeneity of worker groups, which is termed "worker types" in this study, following BLM's terminology.

Formally, the assumption states that for a firm j in class k, the log-wage distribution, for a given time period, can be expressed as:

$$f_k(w_i) = \mathbb{1}\{\hat{k}(i) = k\} \sum_{l=1}^{L} q_k(L(i)) \mathcal{N}(\theta_{kl}) \tag{3.9}$$

where, $f_k(\log(w_i))$ is the log hourly wage mixture of firm class k, when observing worker i. With some abuse of notation, L denotes the number of worker types, $q_k(L(i))$ represents the proportion of workers of type $L(i)$ in class k, and $\mathcal{N}(\theta_{kl})$ is the Gaussian probability density function for type l workers in class k, with θ_{kl} representing the parameters of this distribution. The indicator

function $\{\hat{k}(i) = k\}$ ensures that we consider only the wage distributions of workers assigned to the specific firm class k.

My approach leverages firm clustering to address the dimensionality challenge inherent in firm heterogeneity analyses. By aggregating individual firms into a more manageable set of "firm classes", I circumvent the need to restrict the dataset to a set of connected firms through workers. However, the identification strategy of this methodology still relies on job movements. It shifts, however, the focus from tracking movements between individual firms to observing transitions across firm classes. Therefore, while this mixture model still fundamentally relies on job mobility, it does so at a more aggregated level. In the supplementary material, I perform a clustered AKM regression to show that on average, the residual change in wages for these movers is close to zero, suggesting the movement pattern is not related to the labor market structure itself.

A crucial assumption of this approach is that each worker type, to be estimated in the second step, exhibits a unique pattern in their "cycling" through firm classes as they navigate job changes. These transitional pathways must be sufficiently distinct to allow for clear identification of worker type parameters (Bonhomme, Lamadon, and Manresa, 2019). The robustness of this assumption in my context is based on the substantial number of observations in the dataset, which provides the statistical power necessary to discern these distinct mobility patterns between worker types and firm classes.

The k-means algorithm aims to group firms with similar payment schedules. Formally:

$$\underset{k(1),\dots k(J),H_1,\dots,H_K}{\arg\min} \sum_{j=1}^{J} n_j \int (\hat{F}_j(w) - H_{k(j)}(w))^2 d\mu(w) \tag{3.10}$$

where $\hat{F}_j$ represents the empirical CDF of the log-weekly wages w of firm j, μ is a discrete measurement, supported by a finite grid of ventiles from the population. K, the number of firm classes, is known, while the array $k(1), \dots, k(J)$ represents the partitioning for each firm. H_k represents cluster k's CDF. Finally, n_j is the firm's corresponding workforce size. I perform 1000 repetitions to ensure a global minimum distance estimation.

In simple terms, this procedure minimizes the distance between firms and unobserved classes using as measurements each firm's empirical CDF generated from the ventiles of the observed population log hourly wage distribution. It imposes a weighting parameter to ensure different minimization process for

larger firms. For each biennial sample, I assume that the firm class classification is time-invariant.

I choose $K = 10$ as the baseline number of groups since it minimizes the wage variance within each group. I follow Bonhomme and Manresa (2015) and Bonhomme, Lamadon, and Manresa (2019), where the estimation of firm classes does not affect parameter estimation in the Gaussian mixture step. Nevertheless, in the Appendix , I provide a comprehensive cluster choice analysis using gap statistics to find optimal K-Means clustering estimation. I also provide alternative cluster settings as robustness checks in the discussion section.

3.4.2 Gaussian Mixture Estimation

I assume that observed wages follow a mixture of log-normal distributions, where every "latent" probability distribution represents an interaction of a latent worker "type" with the respective firm class. This approach enables me to reduce the high-dimensional unobserved heterogeneity among individual workers into a manageable set of Gaussian distributions.

I estimate the parameters with the pooled dataset, not observing gender at first. By not accounting for gender at the outset, I ensure that male and female workers assigned to the same distribution are as similar as possible in terms of unobserved characteristics. The idea is that the algorithm will approximate individuals with sufficiently similar unobserved characteristics that spawn the same distribution of wages, regardless of gender. It allows for a more precise comparison of how these latent worker types interact with firm classes without biasing the results by preemptively imposing gender differences.

This approach allows for a more flexible examination of the wage structure assumption in the labor market. By constructing and comparing expected payment levels for each worker type and firm class interaction, I can empirically assess at which extent the additive separability assumption hold, and capture interactions in the market that deviates from this condition. Finally, I can disaggregate these payment levels by gender to measure the differential complementarity effects on wages, providing insights into how worker-firm interactions contribute to gender wage disparities, especially at matches where the separable form is not observed.

Recovering Worker Types

To identify latent worker types, I posit that the wage distribution for each type depends on their associated firm class and follows a log-normal distribution. This approach incorporates potential complementarities characteristic of spe-

cific worker-firm matches. I first, estimate the densities for job movers, and subsequently, I estimate the proportions of stayers using the job mover distributions from the initial period.

I formulate this as a maximum likelihood problem, closely following Bonhomme, Lamadon, and Manresa (2019):

$$\arg\max_{\theta_p,\theta_1,\theta_2} \sum_{i=1}^{N_m} \sum_{k=1}^{K} \sum_{k'=1}^{K} \mathbb{1}\{\hat{k}_{i1} = k\}\mathbb{1}\{\hat{k}_{i2} = k'\} \log\left(\sum_{l=1}^{L} p_{kk'}(l;\theta_p) f^1_{kl}(w_{i1};\theta_1) f^2_{k'l}(w_{i2};\theta_2)\right) \tag{3.11}$$

where N_m denotes the number of job movers, K the number of firm classes, and L the number of worker types (set to 10 for interpretability). The indicator functions $\mathbb{1}\{\hat{k}_{i1} = k\}$ and $\mathbb{1}\{\hat{k}_{i2} = k'\}$ capture the transition of worker i from firm class k to k' between periods 1 and 2. $p_{kk'}(l;\theta_p)$ represents the proportion of type l workers moving from class k to class k', while f^1_{kl} and $f^2_{k'l}$ are log-normal wage distributions for type l workers in classes k and k' in periods 1 and 2, respectively.

Therefore, Equation 3.11 captures the parameters of the conditional distributions of the worker types leveraging the job movers.

For job stayers, I estimate:

$$\arg\max_{\theta_q} \sum_{i=1}^{N_s} \sum_{k=1}^{K} \mathbb{1}\{\hat{k}i1 = k\} \log\left(\sum_{l=1}^{L} q_k(l;\theta_q) f^1_{kl}(w_{i1};\hat{\theta}_1)\right) \tag{3.12}$$

where N_s is the number of stayers, and $q_k(l;\theta_q)$ is the proportion of type l stayers in class k. I leverage the first year parameters for job movers. I employ the Expectation-Maximization (EM) algorithm with 50 repetitions to estimate these parameters.

To recover the most likely worker type for each observation, I utilize the Maximum A Posteriori (MAP) estimation. Formally, for a worker i in firm class k with wage w_i, the probability of belonging to type l is given by:

$$P(l|w_i,k) = \frac{q_k(l;\hat{\theta}q) f_{kl}(w_i;\hat{\theta})}{\sum_{l'=1}^{L} q_k(l';\hat{\theta}_q) f_{kl'}(w_i;\hat{\theta})} \tag{3.13}$$

The worker type is then assigned as:

$$\hat{l}_i = \arg\max_{l} P(l|w_i,k) \tag{3.14}$$

3.5 Estimated Parameters

In this section, I present the estimated parameters for the mixtures, beginning with firm class estimates, followed by the mixture proportions, and concluding with a detailed analysis of the estimated moments disaggregated by gender.

3.5.1 Cluster eCDFs

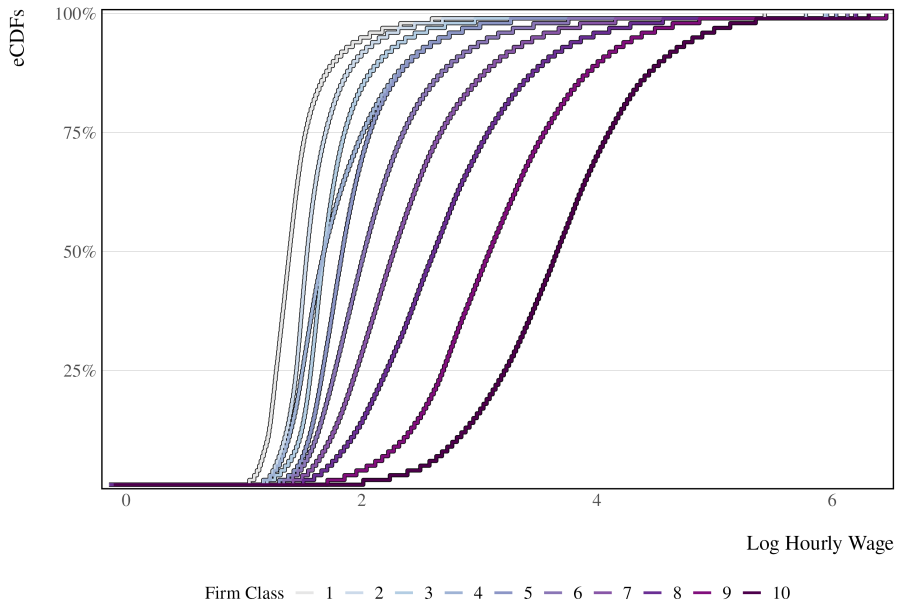
The effectiveness of the algorithm in segregating firms into distinct clusters is evaluated by visualizing the empirical cumulative distribution function of the generated clusters. They are illustrated in Figure 3.1a.

As depicted, the algorithm managed to delineate mostly clear firm classes, grouping firms with similar pay policy, evidenced by the "clear cuts" of each cluster's CDFs, with the exception being firm class 4.

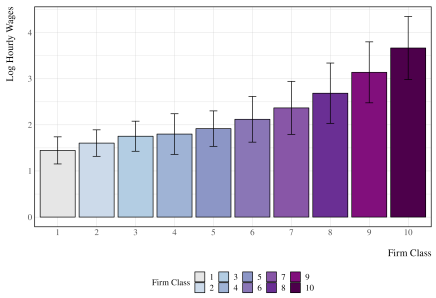
Figure 3.1b provides the moments of their log hourly wage distribution, with the means as the bars and the first standard deviation as the error-bars. For each estimated cluster, not only expected payment increase but also their dispersion when going upward in the firm class ranking. For example, the lowest firm class pays, on average, 1.45 in log hourly wages, with a variance of 0.09, while the highest pays 3.67 with 0.47 in variance.

Figure 3.1c reveals the gender wage gap in means (expressed as $\mathbb{E}[w_{it}^f|k] - \mathbb{E}[w_{it}^m|k]$) as the line plot (the right y axis), and the average size per firm as the bar plot (the left y axis). The expected gender wage gap in means has a tendency to increase when going up in firm class ranking. The lowest paying firms are the most equitable firms in the labor market, with the lowest difference between genders at 11 log-points. The plot also reveals highest-paying firms, which tend to be larger firms[11], exhibiting the largest gender wage disparities, reaching 25 log points. This finding is not entirely unexpected given the substantial variance in wages within firm class 9 or 10. This pattern suggests potential overestimation of the magnitude of firm effects contribution to the gender gap under additive separable models. This overestimation likely stems from the necessary practice of focusing on large firms to ensure sufficient worker mobility within a connected set, while addressing the "double-coincidence" problem of observing both male and female job transitions. However, this approach inadvertently oversamples precisely those firms where gender wage disparities are most pronounced, potentially skewing overall estimates of firm effects on the wage gap.

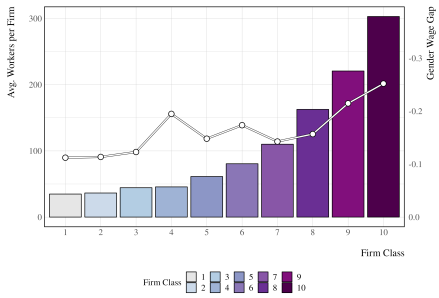
[11] For a full descriptive statistics of firm classes, see Table C.5.2 and C.5.3



(a) Empirical CDF of Firm Classes



(b) Wage (Mean and Variance) Statistics



(c) Gap and Size Statistics

Figure 3.1: (a) Firm Class ECDFs, (b) Firm Class Mean and Variance, and (c) Firm Class Size and Gender Wage Gap Statistics

Note: [1]Firm classes estimated by a k-means clustering algorithm using as measurement their empirical cumulative distribution function supported by the ventiles of the population. [2]The Gender wage gap in means (line in Panel C) is calculated as the female minus male: $\mathbb{E}[w_{it}^f|k] - \mathbb{E}[w_{it}^m|k]$.

3.5.2 Assortative Matching of Estimated Parameters

Figure 3.2 displays the unconditional distribution of workers across firm classes (top row) and worker types (bottom row) for each gender. Both male and fe-

male workers exhibit a concentration of employment in firm class 6, but the proportion is slightly higher for men, with 15 percent of the male workforce in this class compared to 13 percent for women. Additionally, the distribution for men shows a more noticeable skew towards higher-productivity firms. Specifically, 17 percent of men are employed in the top two firm classes (9 and 10), whereas about 14 percent of women are employed in these high-productivity firms. This suggests that men are more likely to be employed in firms that offer higher wage premiums, which may contribute to the observed gender wage gap through the sorting channel.

The differences in distribution become more pronounced when examining worker types. The female distribution is heavily skewed to the left, with nearly 24 percent of women concentrated in worker type 3 versus 17 percent among male workers. In contrast, the male distribution is more evenly spread across worker types, exhibiting a more balanced, albeit still slightly left-skewed, pattern.

In this paper, worker types represent comparable unobserved heterogeneity. Meaning female and male type 3 are individuals where their wages are likely drawn from the same set of Gaussian distributions. The firm class distribution has a more straightforward interpretation, as the proportion of firms with similar payment policies, mirroring patterns of productivity and industry.

When I discuss the gender wage gap decomposition, I hold the distribution of worker types constant since channels of worker type heterogeneity may arise from a multitude of mechanisms in the labor market, such as non-monetary preferences or human capital levels.

Figure 3.3 displays firm classes along the horizontal axis against the stacked conditional proportions of corresponding worker types, separately for female and male workers. These proportions are recovered by grouping types for each male and female sample conditional on each firm class after the *maximum a priori* classification.

Worker types and firm classes are numbered according to expected payment. Therefore, type 10 represents on average the highest paid worker in the data, a proxy for individuals that overall possess high human capital value. The visual representation clearly illustrates an assortative matching pattern, revealing that higher-paying firms predominantly employ higher types of workers for both genders. However, there are notable differences between male and female sorting patterns.

For female workers, there is a strong concentration of lower-type workers in lower-class firms. For instance, in firm class 1, 29 and 23 percent of the workforce comprise of type 1 and type 2 workers, with another 36 percent being type 3 and

4 together. Moving to higher firm classes, this composition shifts dramatically: in class 10, less than 5 percent are type 1 workers, while 15 and 26 percent belongs to type 10 and 9 workers.

On the other hand, male workers shows a slightly different trend. Type 1 and 2 workers comprise together 44 percent of firm class 1 workforce, slightly less concentrated than for females. In the highest firm class, while also presenting negligible proportions of the lowest type, 51 percent of the workforce is comprised of type 10 and 9 workers.

Therefore, while assortative matching is evident for both genders, the patterns reveal some disparity in how men and women with sufficiently similar unobserved heterogeneity are sorted across firm classes. Women appear to face some friction in ascending the firm classification hierarchy, resulting in a more pronounced concentration in lower-tier firms even when their latent productivity (as captured by worker types) is comparable to that of their male counterparts.

Theil Index

To quantitatively assess whether the male distribution in the labor market is slightly more symmetrical compared to the female, I perform a Theil Index calculation to measure the inequality, where the metric is the number of workers per match. I separate the firm classes into low and high classes, where low comprises of firm class 1 to firm class 5, while high comprises of firm class 6 to firm class 10.

The simple Theil Index formula is:

$$T = \frac{1}{M} \sum_{m=1}^{M} \frac{N_m}{\bar{N_m}} \log \left(\frac{N_m}{\bar{N_m}} \right) \tag{3.15}$$

where m is each match in the labor market, M the total number of matches, in this case, 100. N_m is the total number of workers for each match, while $\bar{N_m}$ is the average number of workers per match in the labor market.

The Theil index for the male distribution is 0.43. For the female distribution is 0.48, slightly larger, suggesting that the female distribution of workers in the labor market is more sorted towards the left, concentrated in overall less paying matches.

3.5.3 Payment Schedules

The BLM method not only demonstrates the flexibility to capture assortative matching but also enables researchers to discern the underlying wage structure

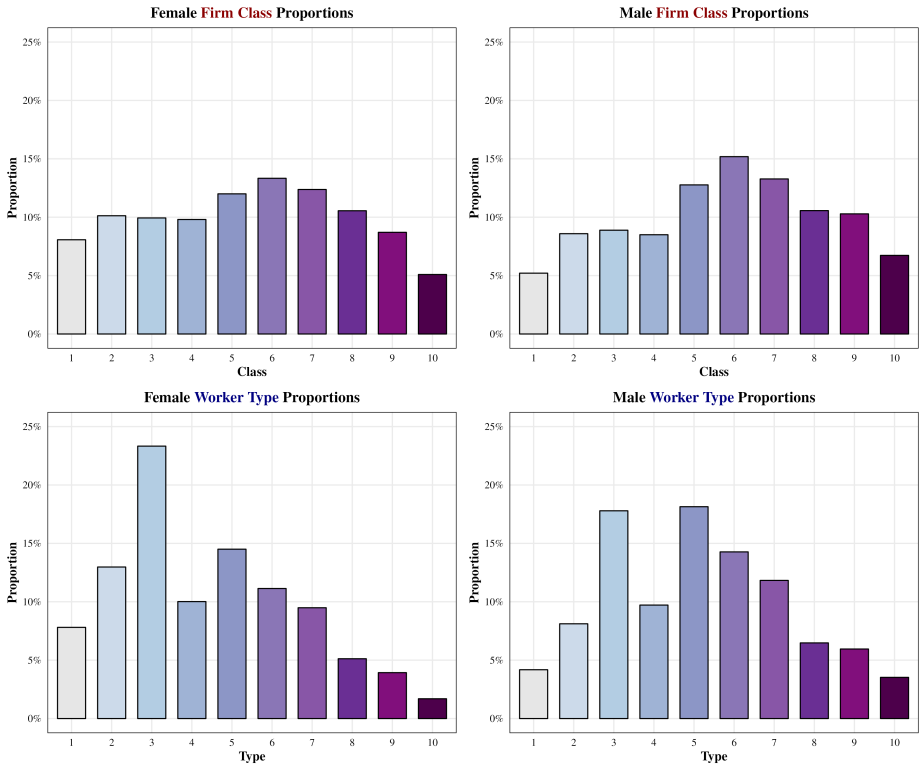


Figure 3.2: Worker Type and Firm Class Unconditional Probabilities per Gender

Note: [1] Firm class estimated using k-means clustering of the cumulative distribution function of payments. Worker types estimated using a Gaussian mixture model where I assume each latent worker type interact with firm classes by drawing wages from a log-normal distribution.

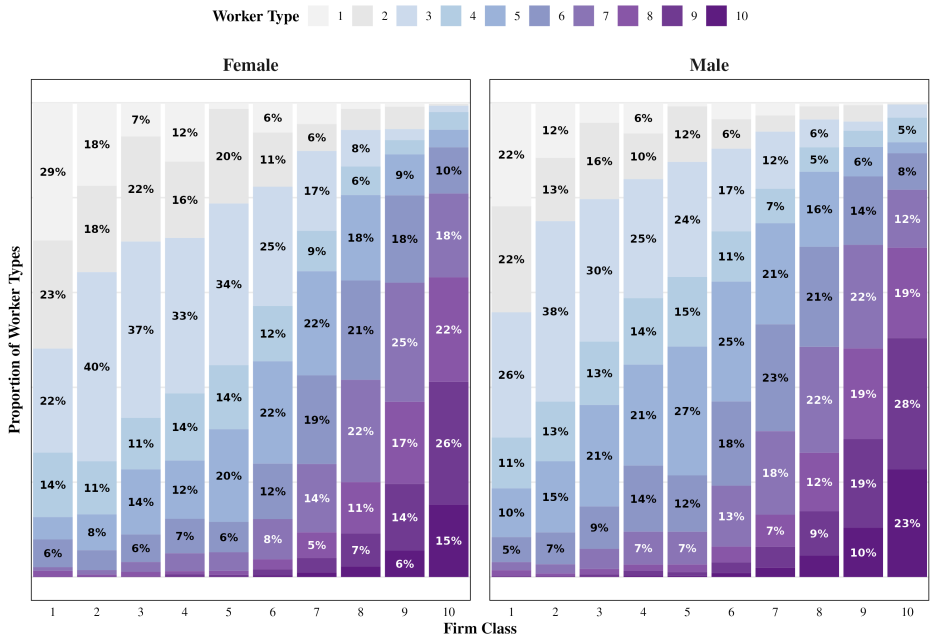


Figure 3.3: Proportion of Estimated Worker Types and Firm Classes

Note: [1]Proportions of worker types recovered using a finite Gaussian mixture of log hourly wages conditional on observed firm classes. Firm classes recovered using k-means clustering algorithm on firm's log hourly wage's CDFs. [2]Worker type membership assigned using a *maximum a priori* estimation. [3]Types and classes ordering based on expected log hourly hours.

arising from firm-worker interactions by assessing whether certain compensation patterns result in wage levels that surpass predictions from additive separable models.

Figure 3.4 presents the payment schedules by firm class and worker type. Panel (a) represents estimated average payments directly under the Gaussian mixture model. Panel (b), on the other hand, is a counterfactual scenario where worker and firms do not yield complementary wage effects in their interactions. I construct this counterfactual by performing an "AKM" two-way fixed effect model such as:

$$w_{it} = \alpha^{g}_{L(i)} + \psi^{g}_{K(i,t)} + \varepsilon_{it} \tag{3.16}$$

where w_{it} is the log hourly wage of worker i in time period t, $\alpha_{L(i)}$ is the fixed effect of worker i's type l, represented under the assignment function $L(i) = l$, $\psi^{g}_{K(i,t)}$ is the fixed effect of firm class k, also represented under an assignment function $K(i,t)$. ε_{it} is the idiosyncratic error term. To make sure I preserve gender disparities, I regress twice for each gender sample.

I introduce a weighting parameter to mitigate the influence of extreme values on my estimation. It leverages the fact that common interactions in the labor market tend to possess small complementarity effects[12]. Consequently, the objective function for this minimization problem can be expressed as:

[12] Figure 3.3 reveals that "extreme complementarity" matches are approximately 5 percent of the total.

$$\min \sum_{i,t} n_{M(k,l)} (w_{it} - \alpha^{g}_{L(i)} + \psi^{g}_{K(i,t)})^2 \tag{3.17}$$

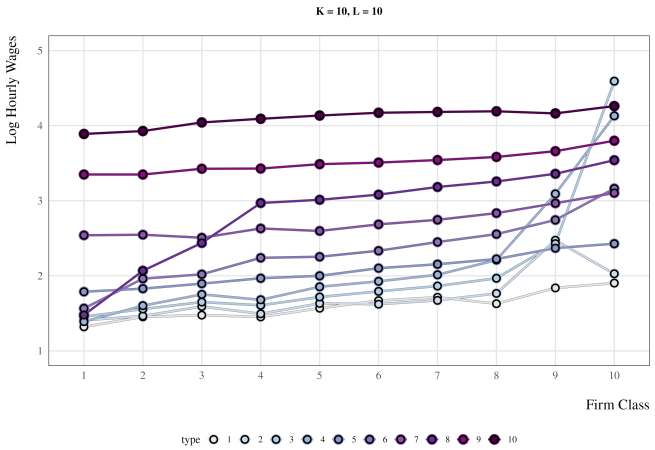
where $n_{M(k,l)}$ represents firm class k and worker type l match's proportion of the number of workers.

Each panel in Figure 3.4 shows each line representing an expected payment "path" of each worker type when hired by a particular firm class.

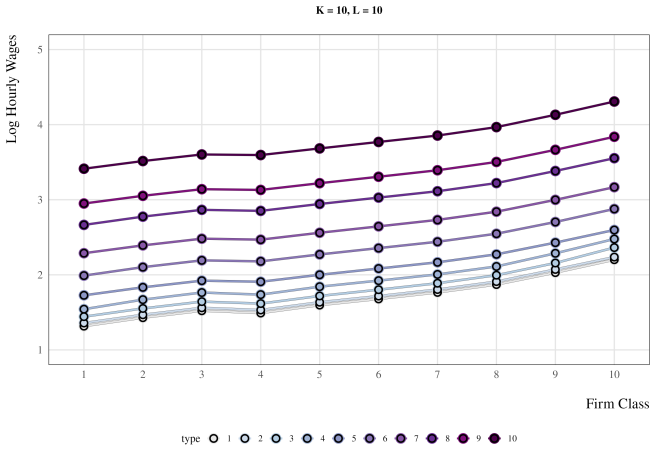
The Gaussian mixture model is able to capture different wage levels that do not necessarily follow a linear trend, as shown in Figure 3.4a.Top firm classes tend to offer substantially higher wage levels to individuals, with particularly pronounced effects for workers in the lower to middle range of the skill distribution. High-earning individuals exhibit remarkable wage stability across firm classes, maintaining their elevated earnings even when matched with low firm classes, with a slight decrease. There is also severe wage compression at the left tail of the distribution. In particular, "worker type 8" experiences severe wage compression if matched with extreme low firm classes such as 1 or 2.

When worker-firm interactions are assumed to be "additive separable", particular interactions are smoothed out as shown by Figure 3.4b, the lines become parallel, which is the quintessential feature of the additive separability assump-

tion: workers and firms contribute to the wage generation function by adding their respective "values". That means worker type 10, for example, if transferred from firm class 10 to 1, should not lose the part of their wage that belongs purely to their components.



(a) Gaussian Mixture Estimated Wages per Match



(b) Additive Separable Prediction of Wages per Match

Figure 3.4: Payment Schedules of worker-firm interactions under Gaussian mixture estimates and predicted linear model.

Note: Panel (a) generated by using estimated means and variances of each Gaussian component of the mixture. Panel (b) generated by running a two-way fixed effect estimation with firm classes and worker types as fixed effects, weighted by the number of workers per each match.

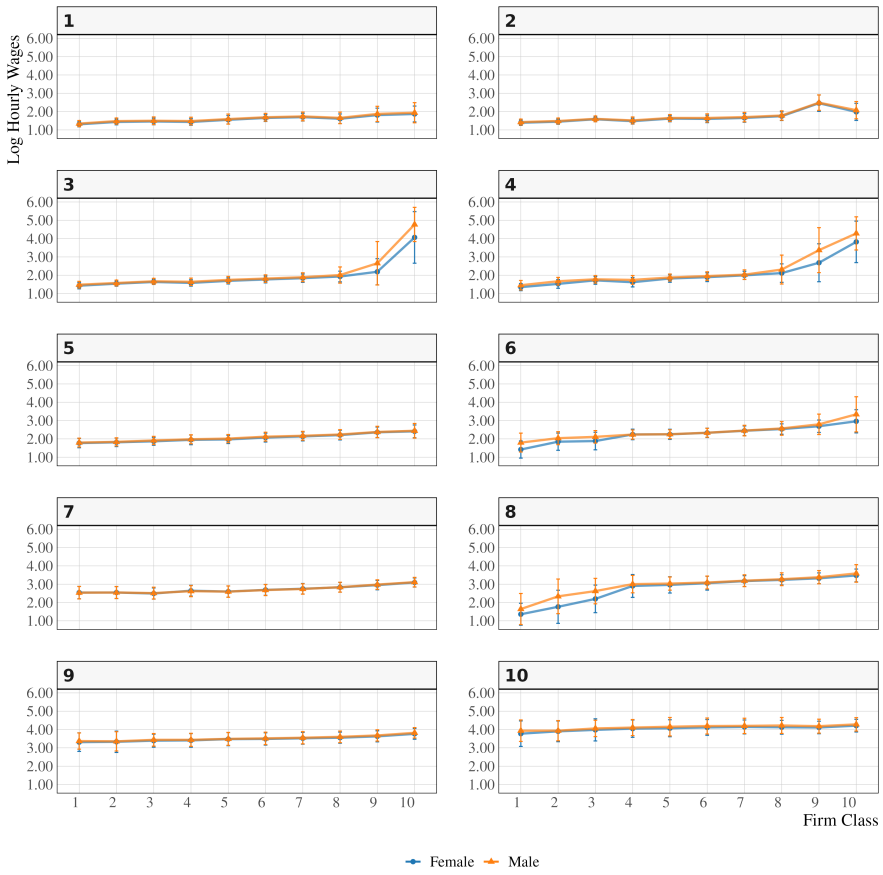


Figure 3.5: Pay Schedules by Gender, Firm Class, and Worker Type

Note: [1]Each panel represents a worker type payment schedule for each firm class, in log hourly wages, grouped by gender. [2]Points indicate the mean log hourly wage, and error bars represent one standard deviation of the estimated wage distribution.

Gender-wise Payment Schedules

A natural question that arises is to what extent these particularities matter for the creation of pay differentials between male and female workers. I construct Figure 3.5 by expanding Figure 3.4a, separating each worker type into a male and a female components.

Complementarity effect arising from those matches implies significantly higher wages for men as compared to women, in particular for the worker types 3, 4, and 8, for which the deviation from a "separable setting" is most extreme. For example, type 3 matched with firm class 10 yields almost 1 log hourly wage gap. Type 4 under the same match yields about 50 log-points in gap.

Type 8, and to a lesser extent type 6, exhibit negative deviations as they approach the lower extreme of firm productivity. When these worker types, characterized by moderate-to-high human capital accumulation, are found in low-productivity firms, a compression effect on wages emerges. In these cases, the expected wage falls below the sum of the expected firm and worker effects. Female workers are more susceptible to these unfavorable matches compared to their male counterparts.

Workers Under Comparative Advantage

To understand better the differences being under a comparative advantage match and otherwise in the labor market, I compare individuals with sufficiently close payments but differing in matches. Specifically, I compare the type 3 and type 10 male and female workers when hired by the class of firm 10. Table 3.3 provide a descriptive statistics of these workers.

The table highlights distinct differences in education, age distribution, and occupations between both types of workers when under firm class 10, male and female. Both groups, regardless of gender, possess a high concentration of college degree individuals, with "type 10 workers" having slightly more. Age is also similar, whereas less than 30 years old female workers are more likely to be found under type 3, while 31-50 are marginally more likely to be found under type 10.

For occupation, while both types display a higher concentration of scientific and liberal arts [13], there is a much higher concentration of managers.

Even with similarities, there is some evidence that individuals in complementarity effect matches might have leading positions and are particularly valuable for firms to employ. Such individuals would be suffering more extreme wage compressions if hired elsewhere, while worker type 10 experiences a more predictable wage path along firm classes.

[13] These categories are generated from the Brazilian Code for Occupation, and tend to have similarities with other codes internationally. "Scientific and Liberal Arts' is a generic code that summarizes economists, engineers, lawyers, professors, among others, whose jobs under normal circumstances require at least a degree from a superior institution of learning or education.

3.6 Discussions: Monte Carlo Simulation and Variance Decomposition

In this section, I introduce a novel decomposition of the gender wage gap that accounts for complementarity effects in the labor market. I decompose the gender wage gap into three distinct components. The first component captures the contribution of complementarity effects, which I isolate by constructing a counterfactual labor market without comparative (dis)advantage matches.

Table 3.3: Workers Under Complementarity and Non-complementarity Matches in Firm Class 10

	Type 3 Workers		Type 10 Workers	
	Female (1)	Male (2)	Female (3)	Male (4)
Education And Age				
Dropout	0.04	0.01	0.00	0.01
High School Graduates	0.12	0.06	0.03	0.04
College	0.84	0.92	0.97	0.95
Age (<30)	0.15	0.06	0.10	0.09
Age 31-50	0.69	0.66	0.74	0.70
Age (≥51)	0.13	0.24	0.14	0.18
Occupation Statistics				
Scientific and Liberal Arts	0.20	0.25	0.36	0.37
Technicians	0.04	0.03	0.08	0.11
Administrative	0.18	0.07	0.19	0.13
Managers	0.52	0.62	0.35	0.36
Traders	0.05	0.01	0.02	0.02
Rural	0.00	0.00	0.00	0.00
Factory	0.01	0.01	0.01	0.01
Mean experience (years)	6.375	7.302	8.417	8.288
Mean Log-Wage	4.308	4.866	4.204	4.270
Variance of Log-Wage	1.836	0.715	0.094	0.116
Worker-years observations	5718	18 401	73 869	157 282
Number of Workers	4895	15 542	44 309	93 073
Fraction of Women	0.24	0.76	0.32	0.68

Notes: [1]Under complementarity matches are type 3 and type 4. Without complementarity is type 10 match, which in wage levels is comparable to matches under complementarity. [2] Education, age, and occupation statistics are fractions that may not necessarily add to one due to rounding.

The remaining two components are inspired by Card, Cardoso, and Kline (2016). The second component, referred to as the "sorting" component, reflects the impact of firm allocation on the wage gap. I calculate this by simulating male and female labor markets where all factors are held constant except for the distribution of firms.

The final component, the "bargaining" component, represents the wage gap contribution arising when equally productive individuals are employed by firms of the same class, but a gender-based differential persists. I isolate this effect through a simulation in which male and female labor markets share identical distributions of workers and firms, while the means and variances of each gender-wise Gaussian distribution remains as observed in the original data.

The "bargaining" and "complementarity" components share certain similarities in nature. The complementarity effect can be viewed as a subset of the bargaining effect in the context of a CCK framework, as both reflect differences in returns for similar individuals within the same firm. However, the Gaussian mixture model allows me to distinguish between these two components, as it identifies labor market matches where wages deviate from the assumption of additive separability. As a result, complementarity effects emerge only in these specific labor market settings, whereas bargaining is more applicable in contexts where the additive separability condition holds.

3.6.1 Monte Carlo Simulations

To setup the Monte Carlo Simulations, I first calculate the realized moments of every worker type and firm class match in the labor market for the male and the female sample. Then I calculate the unconditional probabilities of worker types and firm classes for male and female[14]

To create a separable market, I match workers following a "diagonal pattern in matches" in Figure 3.4a. That means type worker 10 is guaranteed to work in firm class 10 as long as there is a spot available. When firm class 10 job slots are filled, firm class 9 starts hiring the best available, until all jobs are filled with workers. Figure 3.6 shows the resulting conditional probabilities of worker types given firm classes and gender under a separable market.

In a labor market characterized by strictly additive separability, reshuffling matches is expected to have negligible effects on overall wage levels (Graham, Imbens, and Ridder, 2014). Therefore, I leverage this fact, and the fact that "diagonal" matches do not yield large complementarity effects, to construct a labor market that behaves under the additive separability assumption. As a robustness check, I also perform my analysis using the weighted linear regression predicted fixed effects shown in Figure 3.4b.

[14] Means and the standard deviations of each match are shown in Figure 3.5. The unconditional probabilities are shown in Figure 3.2.

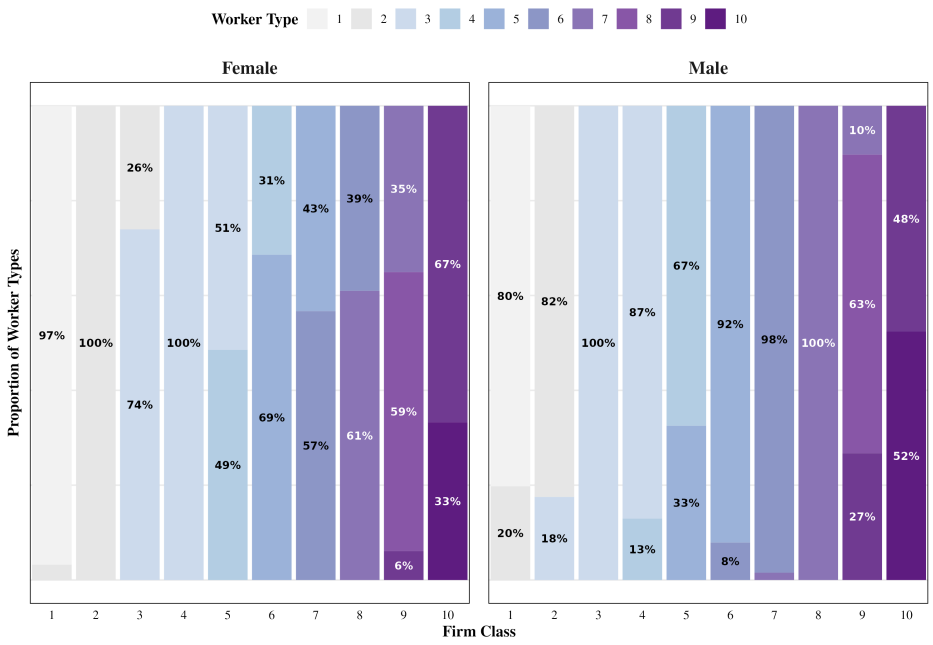


Figure 3.6: Conditional Probabilities of Worker Types Given Firm Classes and Gender, Under a Separable Market

Note: Probabilities calculated by generating a labor market where "top workers" are guaranteed to match with "top firms", until all positions are filled, following Becker (1973)'s principle of assortative matching.

The difference in the gender wage gap between the separable labor market and the original setting reflects the contribution of complementarity effects that are not captured in the separable model. For CCK's "bargaining" and "sorting" components, I conduct simulations within the separable labor market to ensure that these components remain distinct and do not overlap.

Simulation Results

Table 3.4 presents the results of the simulations. The first row displays the observed mean log hourly wages for female and male workers, along with the difference in log points, for the overall dataset, as well as broken down by education cohort and age group. The first column reports the difference in means calculated using the Gaussian distribution, which closely mirrors the observed gender wage gap. The second column reflects the gender wage gap in the counterfactual world where the labor market is additively separable.

The next three columns represent the contributions of the three proposed components. The first column shows the difference between the baseline and the separable market, indicating the contribution of complementarity effects

to the gender wage gap. The "sorting" component reflects the contribution of firm allocation to the wage gap, holding all else constant in the separable market except the distribution of firms. Lastly, the "bargaining" component captures the contribution to the wage gap when all else is held constant in the separable market, except for the observed moments (means and variances) of each Gaussian distribution.

Table 3.4: Gaussian Mixture Decomposition of Gender Wage Gaps

			Contribution to Gender Wage Gap		
Group	Baseline Market Gap (1)	Separable Market Gap (2)	Complementarity Contribution (3)	Sorting Contribution (4)	Bargaining Contribution (5)
All	-0.24	-0.20	-0.04 (16.7%)	-0.09 (37.5%)	-0.02 (8.3%)
Education					
No high-school	-0.30	-0.27	-0.03 (10.0%)	-0.13 (43.3%)	-0.03 (10.0%)
High-school	-0.23	-0.22	-0.01 (4.3%)	-0.08 (34.7%)	-0.02 (8.7%)
College	-0.35	-0.30	-0.05 (14.3%)	-0.16 (45.7%)	-0.04 (11.4%)
Age					
<30	-0.09	-0.08	-0.01 (11.1%)	-0.03 (33.3%)	-0.02 (22.2%)
31-50	-0.30	-0.26	-0.04 (13.3%)	-0.13 (43.3%)	-0.03 (10.0%)
51>	-0.33	-0.28	-0.05 (15.2%)	-0.13 (39.4%)	-0.03 (9.1%)

Notes: [1]All values represent log wage gaps (female - male). Baseline Gap is the observed gap. [2]Separable Market Gap assumes interactions do not yield complementarity effects. [3]Complementarity Contribution is the difference between Baseline and Separable Market gaps. [4]Sorting Contribution is the reduction in gap after equalizing means and variances of worker-firm interactions. [5]Bargaining Contribution is the reduction in the gap after equalizing firm probabilities. [7]Both sorting and bargaining are calculated under a separable market. [6]Numbers in parentheses show the percentage of the Baseline Gap explained by each component.

Overall, the complementarity effect contributes approximately 16 to 17 percent to the gender wage gap, indicating that disparities arising from comparative advantage matches play a significant role in generating wage differences between male and female workers. Labor market allocation accounts for about 37.5 percent, while differences in bargaining without considering complemen-

tarities contribute roughly 8.3 percent. Together, these components explain nearly two-thirds of the gender wage gap. These findings demonstrate that non-separable, two-sided heterogeneity models, such as the Gaussian mixture approach of BLM, more effectively capture the substantial role that firms play in contributing to the gender wage gap, both horizontally as vertically.

The gender wage gap is smaller among individuals with a high school education but reaches its peak among those with college degrees. Furthermore, changes in the gender wage gap are not primarily driven by firm allocations; instead, they are largely explained by complementarity effects, especially for individuals with college degrees. This suggests that women with high levels of human capital, who are in positions of comparative advantage, are particularly susceptible to wage disparities arising from these effects.

Another evidence of the human capital accumulation and complementarity effect positive correlation is the age analysis. While sorting increases its role in wage differentials for individuals older than 30 years, it stabilizes for older than 51, while complementarity effect contribution keeps increasing slightly.

To further expand my study, I present Table 3.5 which shows the results of simulations based on samples of different firm sizes and occupations particularly relevant to this study. I focus on three categories of occupations. The first, "Hotels and Restaurants," includes workers directly involved in the hospitality sector, such as waiters, kitchen staff, and cleaners.

The second category, “economists and engineers”, is self-explanatory. The rationale for selecting these professions lies in the fact that, in Brazil, these fields are highly regulated, meaning that only individuals with the appropriate college degree are legally permitted to practice. This choice offers two key advantages: it controls for college diplomas that are more uniform in their practice than, for instance, medical doctors, but also focusing on degrees that typically lead to higher compensation in the labor market.

Finally, the "Managers" category includes workers in leadership positions. These workers are likely to possess high levels of firm-specific human capital, giving them a strong comparative advantage in the labor market.

If comparative advantage drives complementarity effects on the gender wage gap, then strategic positions in the labor market, those involving valuable human capital accumulation and leadership roles, and the size of the firm, associated with bargaining power, should reveal particularly high levels of these effects.

The firm size panel of Table 3.5 reveals that the gender wage gap increases with firm size, particularly due to the complementarity and the sorting effect, meaning women not only are more likely to be found in low-paying firms when

Table 3.5: Gaussian Mixture Decomposition of Gender Wage Gaps - Firm sizes and occupations

Group	Baseline Market Gap (1)	Separable Market Gap (2)	Contribution to Gender Wage Gap: Complementarity Contribution (3)	Sorting Contribution (4)	Bargaining Contribution (5)
All	-0.24	-0.20	-0.04 (16.7%)	-0.09 (37.5%)	-0.02 (8.3%)
Firm Size					
Firms <10	-0.12	-0.12	0.00 (0.0%)	-0.02 (16.7%)	-0.03 (25.0%)
Firms 10-50	-0.14	-0.13	-0.01 (7.1%)	-0.03 (21.4%)	-0.01 (7.1%)
Firms 51>	-0.21	-0.17	-0.04 (19.0%)	-0.08 (38.1%)	-0.02 (9.5%)
Occupations					
Hotel and Restaurants	-0.12	-0.12	0.00 (0.0%)	-0.04 (33.3%)	-0.03 (25.0%)
Economists and Engineers	-0.39	-0.35	-0.04 (10.3%)	-0.16 (41.0%)	-0.05 (12.8%)
Managers	-0.33	-0.22	-0.11 (33.3%)	-0.08 (24.2%)	-0.03 (9.1%)

Notes: [1]All values represent log wage gaps (female - male). Baseline Gap is the observed gap. [2]Separable Market Gap assumes interactions do not yield complementarity effects. [3]Complementarity Contribution is the difference between Baseline and Separable Market gaps. [4]Sorting Contribution is the reduction in gap after equalizing means and variances of worker-firm interactions. [5]Bargaining Contribution is the reduction in the gap after equalizing firm probabilities. [7]Both sorting and bargaining are calculated under a separable market. [6]Numbers in parentheses show the percentage of the Baseline Gap explained by each component.

controlling for larger firms, but also in positions where men are receiving much higher complementarity compensations.

The last three rows of Table 3.5 represent the results of the simulation of different occupations. Hotel and restaurants are typically occupations assumed in the literature to possess high turnover rate and zero firm premium in wages Card, Cardoso, and Kline (2016) and Casarico and Lattanzio (2024). Therefore, these occupations are expected to have negligible comparative advantage effects. Accordingly, my results suggest that the hotel and restaurants labor market is mostly governed by the additive separable assumption, given that the simulated separable market yielded the exact same wage differentials as the baseline market gap[15], confirming that there is no complementarity effect. However, as the expected human capital accumulation is increased, the gap increases. While a considerable portion of the gap is due to firm allocations for economists and engineers, the complementarity contribution represents 10 percent of the gender wage gap.

[15] For the male and female wage levels for all simulations, refer to Table **??**

For managers, the distance in wage differentials between the baseline market and the separable market is the largest, with the complementarity contribution accounting for about 33 percent of the gap. Moreover, the sorting contribution drastically reduces, from 16 to 8 log-points, falling from 41 percent in contribution to 24 percent.

Under an additive separable model, such as linear regression, the results would suggest that labor market allocations are the primary drivers of the gender wage gap among managers[16]. However, in a non-separable model, I can identify that a substantial portion of the previously unexplained differential is due to specific labor market matches that generate complementarity effects. Because additive separable models assume constant returns to unobserved heterogeneity of workers and firms, these contributions are difficult to capture accurately.

[16] In another example, Card, Cardoso, and Kline (2016) found that firm-related factors contributed approximately 4 percent to the gender wage gap after controlling for managers.

3.6.2 Robustness Checks

I perform a series of exercises to show my results are not sensitive to particular choice of parameters. While keeping the optimal number of firm classes according to the gap statistics ($K = 10$), I vary the number of worker types, which are the number of Gaussians observed in each firm class. I test with $L = 6$, $L = 10$, and $L = 12$. I also provide an alternative simulation of the separable market where I use a weighted ordinary least squares with firm classes and worker types as fixed effects. Following Equation 3.17, the weighted parameter is the fraction of workers of each worker type-firm class match. For the weighted OLS, I maintain $L = 10$.

For the alternative number of worker types, results were consistent across all specifications, with the exception of $L = 12$, that seemed to underestimate complementarity effects, putting slightly more contributions to the sorting and the bargaining contribution. The alternative separable market simulation yielded virtually the same estimates as the original, with a more conservative estimation of the complementarity effects contribution to the gender wage gap.

Despite some differences in estimation, overall, the results indicate that my measurements are not driven by errors arising from the Gaussian mixture estimation, local maxima, or a particular setting.

z

3.6.3 Variance Decomposition

A variance decomposition of log wages in works related to Abowd, Kramarz, and Margolis (1999) (AKM models). It decomposes the variance of log wages into five distinct components: (1) the contribution of worker fixed effects, (2) the contribution of firm fixed effects, (3) twice the covariance between worker and firm effects, (4) the variance of time-varying covariates and their associated covariances, typically captured by period dummies interacted with time-varying human capital indicators, and (5) the residual variance.

The AKM model, however, tends to negatively correlate worker and firm effects (Andrews, et al., 2008), implying a downward bias estimate for assortative matching. Bonhomme, Lamadon, and Manresa (2019) proposed using the dimension reduction technique relying in the Gaussian mixture model to mitigate the bias.

Card, Cardoso, and Kline (2016) found that approximately 10 percent of wage variance can be attributed to assortative matching for both male and female workers. In this section, I apply the framework of Bonhomme, Lamadon, and Manresa (2019) to examine the extent to which assortative matching may be underestimated in the wage variance decomposition used in AKM gender analysis.

I compare three models. The first specification uses clustered firm and individual worker identifiers, related to Bonhomme and Manresa (2015), this clustering approach allows the researcher to maintain the linear assumption but reduces the negative bias in assortative matching. I name this model "clustered AKM", or C-AKM, in which fixed effects for firms are now the firm classes[17]. In the second setting I employ the full BLM approach by leveraging both worker types and firm classes.

Finally, I test the variance decomposition analysis under a classical CCK approach which uses individual firm and worker identifiers as fixed effects. It re-

[17] See Appendix Section C.3 and C.4 for discussions on the bias (also often dubbed "limited mobility bias" in the literature) and using the clustered AKM method to perform a KOB decomposition on the gender wage gap.

quires the largest dual connected set of firms and bias correction. In this study, I use the bootstrapping approximation of Azkarate-Askasua and Zerecero (2023) to correct the assortative matching bias.

Formally, the regression setting is:

$$w_{it} = \Phi^g_{K(i,t)} + \Lambda^g_{L(i)} + x'_{it}\beta + \varepsilon_{it} \tag{3.18}$$

And the variance decomposition can be formally stated as:

$$\begin{aligned} \underbrace{\text{Var}(w_{it})}_{\text{Log Hourly Wage Variance}} = & \underbrace{\text{Var}(\Phi^g_{K(i,t)})}_{\text{Firm Class Variance}} + \underbrace{\text{Var}(\Lambda^g_{L(i)})}_{\text{Worker Type Variance}} + \underbrace{\text{Var}(\varepsilon_{it})}_{\text{Residual Variance}} \\ & + \underbrace{\text{Var}(x'_{it}\beta) + 2 \cdot \text{Cov}(\Lambda^g_{L(i)}, x'_{it}\beta) + 2 \cdot \text{Cov}(\Phi^g_{K(i,t)}, x'_{it}\beta)}_{\text{Time-Varying Covariates Variance and Associated Covariances}} \\ & + \underbrace{2 \cdot \text{Cov}(\Phi^g_{K(i,t)}, \Lambda^g_{L(i)})}_{\text{Worker Type and Firm Class Covariance}} \end{aligned} \tag{3.19}$$

where w_{it} represents the logarithmic hourly wage of worker i in period t, decomposed as follows: $\Phi^g_{K(i,t)}$ represents the firm effects, where $K(i,t)$ is the assignment function. Here, K can denote either firm classes or individual firm identifiers. The term $\Lambda^g_{L(i)}$ represents individual worker or worker type effects, where $L(i)$ can refer to either a worker type or an individual identifier. Time-varying covariates are represented by $x'_{it}\beta$, while gender heterogeneity is accounted for by the superscript g. Finally, ε_{it} denotes the idiosyncratic error term.

The covariance between worker and firm effects is of particular interest in understanding the dynamics of assortative matching and its impact on the gender wage gap. This component can be potentially underestimated due to the limited mobility bias. I test three variance decompositions from three different settings. The C-AKM model, by clustering firms, potentially reduces the noise in firm effect estimates, allowing for a more stable estimation of the worker-firm covariance, however, it still relies on individual fixed effects. The CCK model under the bootstrapping correction provides a "lower bound" of these estimates, given that the bootstrapping correction is an approximation, not a total mitigation. The BLM model, employing both worker and firm clusters, provides a framework that effectively circumvents the limited mobility bias by coarsening job movements in the dataset at the cost of noisier results.

Table 3.6 presents the variance decomposition results. The first two columns show the results for the BLM decomposition, columns (3) and (4), the clustered

firm AKM methodology. Finally, columns (5) and (6) represents the results for the classical AKM approach from CCK.

Table 3.6: Variance Decomposition of log hourly Wages

	BLM		Clustered AKM		CCK	
	Female (1)	Male (2)	Female (3)	Male (4)	Female (5)	Male (6)
Var(log hourly wage)	0.529	0.650	0.529	0.650	0.642	0.797
Panel A: ***Variance Estimates***						
Firm effects	0.046	0.066	0.020	0.023	0.042	0.044
Worker effects	0.367	0.357	0.377	0.467	0.479	0.626
Time-varying covariates	0.008	0.011	0.009	0.011	0.006	0.008
Cov(Worker, Firm)	0.137	0.122	0.114	0.136	0.106	0.108
Residual	0.062	0.094	0.010	0.012	0.009	0.011
Panel B: ***Share of Total Variance (%)***						
Firm effects	8.8	10.2	3.7	3.6	6.6	5.5
Worker effects	55.2	54.9	71.2	71.9	74.6	78.5
Time-varying covariates	1.6	1.6	1.6	1.7	1.0	1.1
Cov(Worker, Firm)	26.0	18.8	21.5	21.0	16.5	13.6
Residual	11.6	14.5	1.9	1.8	1.4	1.3

Notes: [1]AKM, BLM, and CCK stand for Abowd, Kramarz, and Margolis (1999), Bonhomme, Lamadon, and Manresa (2019), and Card, Cardoso, and Kline (2016), respectively. [2]Clustered AKM represents firm clustered using a kmeans algorithm and individual worker identifiers as parameters. [3]Panel A showcases the magnitude of estimated variance components, while Panel B presents these components as percentages of the total log hourly wage variance. [4]Results are a weighted average based on the six biennial samples' number of observations.

The first row presents the total variance of log hourly wages by gender. Both the C-AKM and BLM models yield similar magnitudes, as they utilize the full set of worker observations. In contrast, the CCK model relies on the connected set of firms through job movers, which tends to overrepresent larger firms, resulting in higher wage variance estimates.

Male firm effect contribution to the wage variance ranges from 5 percent in the male sample under CCK, to 10 percent under BLM. On the other hand, female firm effect contribution ranges from 3.6 in the clustered AKM model, to 8.8 percent under BLM.

Although AKM models attribute the largest portion of wage variance to worker effects (over 70 percent) for both genders, in the BLM approach worker

effects account for a smaller, though still significant, share of wage variance: 41.3 percent for women and 32.1 percent for men.

The reduction in worker effect contribution can be explained by the larger worker-firm covariance contribution, at 26.0 and 18.8 percent of the total variance for women and men. Compared to the "AKM" model, the BLM and the C-AKM model were more effective in capturing assortative matching effects.

Worker-firm covariance is consistently higher for women across all specifications, suggesting that assortative matching is indeed a meaningful contributor to wage dispersion in the labor market, especially for female workers. My results are particularly relevant in the context of recent discussions on the rise of assortativity in labor markets, as highlighted by Song, et al. (2019), who documented the increasing trend of assortative matching in the United States. If men are systematically more likely to find high-paying matches in the labor market, while women are concentrated in lower-paying positions, this could exacerbate the gender wage gap.

3.7 Conclusion

Additive separable models are unable to capture labor market interactions that generate wages based on comparative advantage, where the match between firms and workers results in compensation that exceeds (or are less than) the simple sum of individual worker and firm contributions.

In this paper, I deviate from linear additive models. I use a linked employer-employee dataset covering all firms and workers from São Paulo, Brazil (2010-2017), to apply the two-sided unobserved heterogeneity framework introduced by Bonhomme, Lamadon, and Manresa (2019). This approach allows me to investigate the contribution of specific worker-firm interactions to the gender wage gap, assuming each interaction generates wages drawn from log-normal distributions. This method allows me to capture the complementarity effects arising from particular worker-firm assortative matching.

Employing Monte Carlo Simulations, I propose a novel decomposition of the gender wage gap into three components. Following Card, Cardoso, and Kline (2016), the sorting component, representing labor market allocations, and bargaining component, representing differences in negotiation of equally productive workers under the same firm. The third component, the complementarity component, is a special case of "bargaining", however, under these matches wage levels do not correspond to the predicted from additive separable model.

I find a positive relationship between human capital and these complementarity effects. They are more pronounced for male workers compared to female workers, accounting for approximately 17 percent of the overall gender wage gap. This contribution go as far as a third of the gender wage gap for individuals in leadership positions. Controlling for occupations that generally require lower levels of human capital, such as occupations related to the hospitality sector, yielded negligible results.

I also find that these interactions are more present at the tails of the wage distribution where larger firms operate and wage dispersion is higher. For firms larger than 50 employees, about a fourth of the gender wage gap is explained by these complementarity effects.

My study demonstrates that firms and the broader labor market structure play a more significant role in shaping the gender wage gap than previously recognized. I demonstrate that firms not only provide varying wage premiums but also evaluate human capital and other worker characteristics in heterogeneous ways. This differential valuation of worker attributes across firms contributes substantially to gender-based wage disparities.

The pronounced complementarity effects observed in managerial positions suggest that policies aimed at increasing transparency in the labor market and promoting key leadership roles among female workers are essential for reducing gender wage disparities.

Future research could leverage on the increased availability of linked employer-employee data and computational power. They could extend the analysis by providing a dynamic framework and exploring how worker-firm interactions evolve over time in response to earnings shocks and their influence on mobility decisions. Incorporating collective bargaining data would further enhance our understanding of the non-monetary factors that shape gender-specific sorting patterns. Expanding this methodological approach to different countries could provide valuable cross-national insights into the extent to which gender wage gaps are driven by universal factors or are shaped by specific institutional and cultural contexts.

Appendix A

Additional Content for Chapter 1

A.1 Theoretical Framework

The core of this analysis rests on two established frameworks of economics, namely the productivity framework and the Rosen-Roback model. These two perspectives allow us to explore the multi-faceted impact of the Mariana disaster, from productivity shocks to spatial equilibrium, and provide us with an understanding of the economic implications of the event. Through these lenses, I can illustrate how similar environmental catastrophes affect the economy through individual firms' decision-making and individual decision-making processes.

A.1.1 Productivity Framework

The productivity framework forms the foundation of our understanding of a firm's decision-making process, which in turn impacts overall economic output. A widely utilized tool in this context is the Cobb-Douglas production function, which allows us to model the relationship between capital, labor, and output. In its most basic form, a firm's optimization problem can be expressed as follows:

$$\max_{K,L} \quad Y - rK - wL \qquad \text{s.t.} \quad Y = AK^{\alpha}L^{1-\alpha} \tag{A.1}$$

Here, Y represents total output, A is the Total Factor of Production (TFP), K stands for the capital stock, L denotes the labor force, and α is the output elasticity of capital. The price of labor, or wages (w), is determined by the firm's optimization of resource allocation. By solving the maximization problem, we get the following First Order Condition:

$$w = (1 - \alpha)A \left(\frac{K}{L} \right)^{\alpha} \tag{A.2}$$

From this, we can observe that the wage level is directly proportional to the level of capital, given $\alpha > 0$. Within the scope of this research, water is treated as a component of productivity, in other words, capital. Consequently, the exogenous shock caused by the dam rupture, which resulted in a decrease in available fresh water and related resources, is expected to exert downward pressure on wages.

A.1.2 Rosen-Roback Model

The Rosen-Roback model is instrumental in understanding the spatial equilibrium across regions or cities, achieved when workers and firms are indifferent between locations. This balance comes from the interplay between wages, rents, amenities, and transportation costs. Specifically, individuals maximize their utility derived from these factors when considering different locations. For example, an individual i will prefer to stay in location a rather than moving to location b if:

$$U_i(W_a) + U_i(A_a) - U_i(R_a) > U_i(W_b) + U_i(A_b) - U_i(R_b) - U(T_{ab}) \tag{A.3}$$

In this equation, W_a, R_a, and A_a represent the wage, rental cost, and perceived level of amenities, respectively, at location a, and similarly, W_b, R_b, and A_b for location b. T_{ab} stands for the transportation cost of moving from a to b, and U_i is a quasi-concave and monotonic utility function. This model predicts that a decrease in perceived amenities could lead individuals to bear transportation costs to move away, prompting firms and landlords to react by increasing wages or decreasing rents, respectively, to retain the population and achieve a new equilibrium.

In the aftermath of the Mariana disaster, it can be hypothesized that office workers and other individuals who do not rely on water for their productivity but perceive the river as an amenity might choose to relocate unless compensated through increased wages or decreased housing costs.

By applying these two theoretical frameworks, I can capture both the direct and indirect impacts of the Mariana disaster on local productivity and spatial equilibrium, providing a comprehensive picture of the economic consequences of such environmental shocks.

A.2 Samples' Summary Statistics Tables

Tables A.1, A.2, and A.3 present summary statistics for the three samples used in this study: the Mariana region sample, known as the Quadrilátero Ferrífero, the extended Minas Gerais sample, which is the Water Basin's analysis sample, and the Espírito Santo sample, used for measuring the impacts on the coastal region. These statistics provide a snapshot of the demographic and economic characteristics of the control and treatment groups in each sample in 2015, right before the disaster.

Table A.1 provides summary statistics for the Mariana region sample. The mean age and wage in the treatment group were slightly lower than in the control group. The treatment group also had a higher percentage of Black, Other, and Pardo individuals and a lower percentage of White individuals. The gender distribution was more balanced in the treatment group, and the education level was slightly lower. Notably, it shows the presence of mining operations in the area, with at least 12 percent of the labor market being composed of mining firms.

Table A.2 presents summary statistics for the extended Minas Gerais sample. The mean age was slightly lower in the treatment group, while the mean wage was slightly higher. The treatment group had a higher percentage of Pardo individuals and a lower percentage of White individuals. The gender distribution was almost identical in the control and treatment groups.

Table A.3 provides summary statistics for the Espírito Santo sample. The mean age and wage were slightly higher in the treatment group. The treatment group also had a higher percentage of Black, Other, and Pardo individuals, and a lower percentage of White individuals. The gender distribution was more skewed towards males in the treatment group, and the education level was slightly lower.

In both the extended Minas Gerais and Espírito Santo samples, the data reveals that mining operations are virtually non-existent. This observation is crucial for disentangling the effects of the dam rupture from the impacts of the ensuing contamination. The evidence strongly suggests that the negative effects observed in the extended Minas Gerais sample are attributable to contamination rather than disruption of mining operations.

Table A.1: Mariana Region Sample Summary Statistics

			Control	Treatment
	Age	Mean	34.80	34.52
	Wage	Mean	2135.30	1845.28
Race	Black	Percent	7.80	10.44
	Other	Percent	10.11	14.48
	Pardo	Percent	46.05	44.81
	White	Percent	36.05	30.27
Gender	Female	Percent	33.97	38.81
	Male	Percent	66.03	61.19
Education	College Education	Percent	9.62	7.96
	High School Education	Percent	57.55	54.40
	No High School Diploma	Percent	32.83	37.65
Industry	Construction	Percent	18.95	13.14
	Financial Services	Percent	0.94	0.82
	Mining	Percent	17.59	12.33
	Public Administration	Percent	1.38	0.91
	Rural Activities	Percent	1.46	1.93
	Tourism	Percent	5.79	6.76
	All	N	146 719	12 389

Table A.2: Extended Minas Gerais Sample Summary Statistics

			Control	Treatment
	Age	Mean	35.13	34.83
	Wage	Mean	1518.98	1523.94
Race	Black	Percent	7.02	5.97
	Other	Percent	8.20	6.10
	Pardo	Percent	20.58	40.59
	White	Percent	64.20	47.35
Gender	Female	Percent	37.78	37.67
	Male	Percent	62.22	62.33
Education	College Education	Percent	7.55	7.12
	High School Education	Percent	47.48	57.96
	No High School Diploma	Percent	44.97	34.93
Industry	Construction	Percent	7.09	10.32
	Financial Services	Percent	1.27	1.54
	Mining	Percent	0.24	0.00
	Public Administration	Percent	0.74	0.03
	Rural Activities	Percent	16.17	4.35
	Tourism	Percent	4.79	5.42
	All	N	1 586 451	235 242

Table A.3: Espírito Santo Sample Summary Statistics

			Control	Treatment
	Age	Mean	34.81	34.97
	Wage	Mean	1503.01	1769.79
Race	Black	Percent	6.66	7.57
	Other	Percent	4.57	5.08
	Pardo	Percent	47.74	56.19
	White	Percent	41.03	31.16
Gender	Female	Percent	43.35	33.99
	Male	Percent	56.65	66.01
Education	College Education	Percent	7.34	6.86
	High School Education	Percent	59.38	54.62
	No High School Diploma	Percent	33.28	38.52
Industry	Construction	Percent	8.98	12.30
	Financial Services	Percent	0.88	0.77
	Mining	Percent	0.76	0.00
	Public Administration	Percent	0.46	0.43
	Rural Activities	Percent	1.49	6.31
	Tourism	Percent	7.48	4.64
	All	N	169 768	329 146

Appendix B

Additional Content for Chapter 2

B.1 Foreign Presence Outside RAIS

Even though RAIS provides a good picture of Venezuelans entering Brazil and exclusively staying in Roraima, it only counts Venezuelans in the formal labor market. Venezuelans may be crossing the border and going through, staying in other states outside the formal labor market, in refugee camps, or working informally. The Federal Police data shows that around 41 thousand individuals crossed Roraima and did not return to Venezuela. However, we do not observe either in RAIS or the Federal Police data whether they stayed in Roraima or moved around.

It would jeopardize our identification strategy if the control states hosted many Venezuelan refugees outside the formal labor market. To show that it is not the case, we rely on the refugee application data from the Brazilian National Committee for Refugees (CONARE).

Foreigners in Brazil can be registered as refugees to get benefits such as obtaining the individual taxpayer registration number (CPF), accessing health and education services, and opening a bank account, among others. A potential refugee must submit its recognition to, and then analyzed by, CONARE. The committee then decides whether they are recognized as a refugee. If rejected, they can appeal. Since refugee applications only exist conditional on the presence of a forcibly displaced population, we believe the data is an adequate proxy for Venezuelans not observed in the formal labor market.

Variables included in CONARE are the nationality of the applicants, the reason for leaving their country, the date when the application was submitted,

the municipality and the state where the application was submitted, and the date when CONARE made the decision.

Table B.1 shows the cumulative number of refugee applications by treatment status and year between 2011 and 2020. There were 56,984 refugee requests in Roraima in 10 years, with the first application submitted in 2015. If we consider our period of interest, 2014-2017, more than 10 thousand individuals requested refugee status, with virtually zero applications found in control states. If immigrants are moving across treated and control states, the likelihood of a Venezuelan applying for refugee status in another location rather than Roraima would significantly increase. Accordingly, we do not see this behavior in the data.

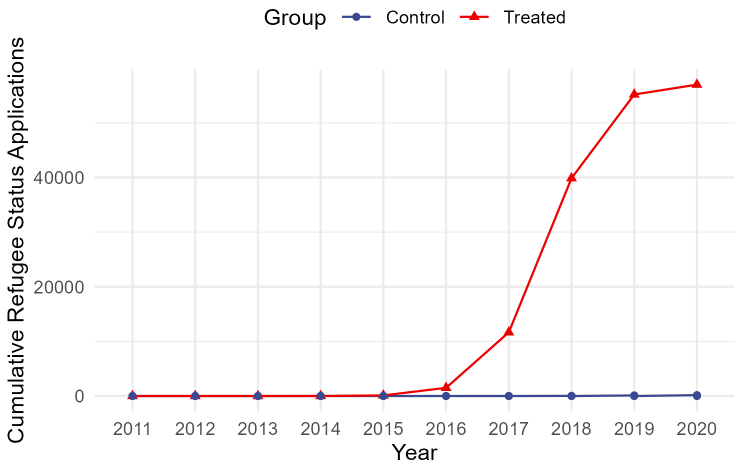


Figure B.1: Cumulative refugee requests by year and treatment status

Another consideration is that only a quarter of those who entered Roraima and stayed in 2017 applied for refugee status. If we combine RAIS, around ten percent of these applicants found a job in the formal labor market. Suppose we assume an individual seeks refugee status to apply for employment and proper residence. In that case, it is a safe inference that a significant fraction of Venezuelans were actively in seek of employment, not necessarily in the formal labor market, with some of these accepting offers in the informal labor market.

B.2 Synthetic Control Methods Results

Concerns may arise in our estimation strategy of using the doubly robust difference-in-differences method given the nature of our natural experiment with only one treatment state (Roraima) if the control group is chosen incorrectly. In the main analysis, we compare the treatment state with carefully chosen control states in Brazil with states with socioeconomic and sociodemographic characteristics comparable to those of the treatment state. We also showed pre-treatment parallel trends in our outcome variable. However, it may lead to endogeneity bias if the control groups are characteristically different from the treatment state.

A popular way of estimating the causal effects of the treatment in the case of a singular or a few treatment states is by using the Synthetic Control Method (SCM). Using SCM with the northern Brazilian states as our donor pool for the synthetic Roraima, we reconfirm the validity of our choice of control states in the main analysis.

Consider the case with $j+1$ states where $j = 0, ..., J$ where $j = 0$ is the only treatment unit, in our case, the state of Roraima. W is the $j \times 1$ vector of non-negative weights for the donor states that comprise synthetic Roraima such that $\sum_J^1 w_j = 1$ and $0 \leq w_j \leq 1$. X_1 is a $K \times 1$ vector of matching variables used to construct our synthetic control group. In our case, our main primary outcome variables, the wages, act as the only matching variable. X_0 is a $K \times J$ matrix for all control states J with the same predictors as Roraima's pretreatment wage trends. The SCM selects the vector of weights $W*$ that minimizes the difference between X_1 and $X_0 W$ such that:

$$W^* = \arg\min_w (X_1 - X_0 W)'(X_1 - X_0 W) \tag{B.1}$$

$$\text{s.t.} \sum_J^{j-1} w_j = 1, 0 \leq w_j \leq 1 \tag{B.2}$$

The optimal W is chosen by an optimization process that minimizes the mean-squared prediction error (MSPE) in the pre-treatment period. Following the construction of weights, we estimate the effects of the wage effects of the Venezuelan refugee crisis in Roraima, Brazil using the formula: $(Y_{Post}^{RR} - Y_{Pre}^{RR}) - (Y_{Post}^{SRR} - Y_{Pre}^{SRR})$, where Y_{Pre}^{RR} and Y_{Post}^{RR} are the average logged wage in Roraima in pre- and post-treatment periods, and Y_{Pre}^{SRR} and Y_{Post}^{SRR} are the average logged wage in synthetic Roraima in pre- and post-treatment periods respectively.

Due to the geographical proximity and sociodemographic similarities with Roraima, our donor pool consists of northern Brazilian states. Figure B.2 il-

lustrates the weights we received from the optimization procedure. The entire weight is shared between Acre (0.80) and Amapa (0.20), which confirms our initial decision to use these states as the control states in our main analysis.

Figure B.3 shows the wage results from the SCM analysis. The figure shows perfect parallel trends between Roraima and synthetic Roraima until 2013 and a few years following. The trends start to diverge in 2016 where the wages for synthetic Roraima have a reduction in slope but Roraima's slope reduction is less steep. In the year 2018, we see an almost 0.20 percentage point difference between Roraima and synthetic Roraima. This exercise confirms the results from our primary analysis using difference-in-differences.

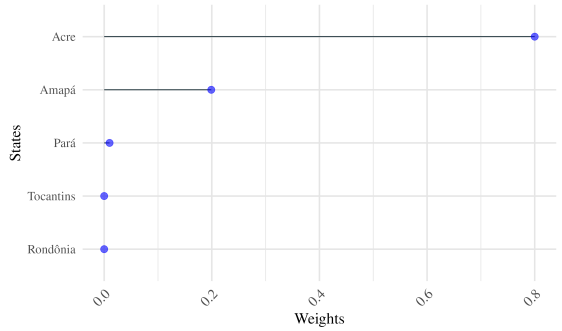


Figure B.2: Weights for Donor States in the Northern Brazilian Region in the Synthetic Control Method

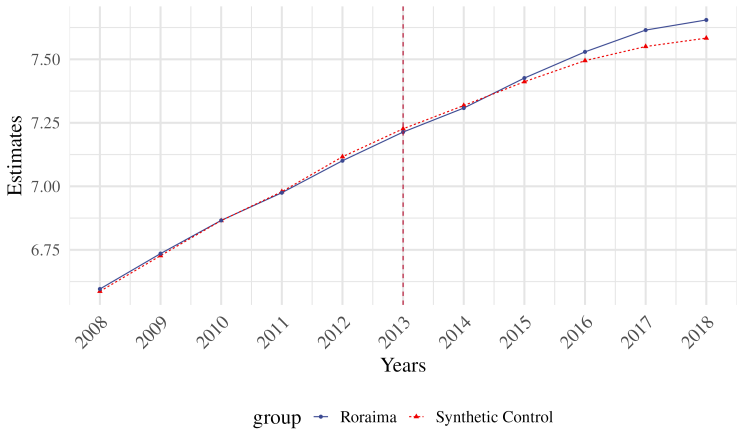


Figure B.3: Effects of the Venezuelan Refugee Crisis in the Brazilian Labor Market using the Synthetic Control Methods

B.3 Synthetic Difference-in-Differences

In this section, we provide results using the Synthetic Difference-in-differences (SDID), approach from Arkhangelsky, et al., 2021, a twist from the classical SCM that relaxes its restrictions by imposing only slope-fitting of the synthetic control and the treatment group.

Figure B.5 shows the optimal donors for the SDID algorithm. It confirms that Amapá and Acre are the most suitable donors for the experiment, with a combined weight value of approximately 65 percent.

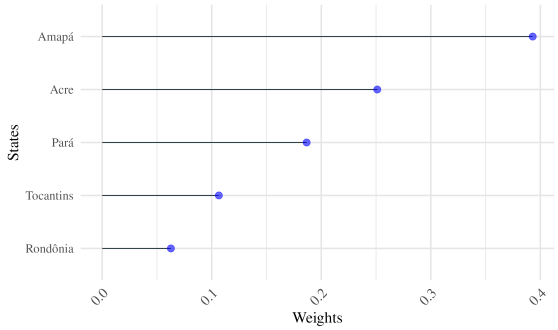


Figure B.4: Weights for Donor States in the Northern Brazilian Region in the Synthetic Difference-in-Differences Method

Figure B.5 shows the treatment group and synthetic Roraima trends using the SDID framework. After 2013, Roraima was able to maintain higher wage levels compared to the generated synthetic Roraima, confirming our main results using a straightforward difference-in-differences approach.

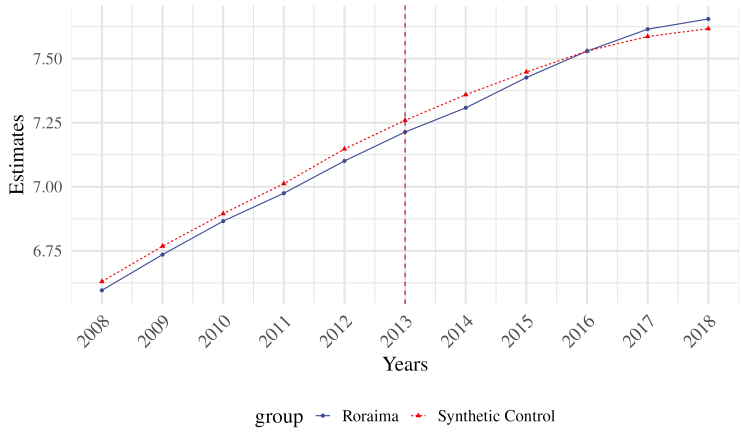


Figure B.5: Effects of the Venezuelan Refugee Crisis in the Brazilian Labor Market using the Synthetic Difference-in-Differences Methods

B.4 Other Figures

Figure B.6: Industries where Venezuelan migrants worked in 2018 (as a percentage of total Venezuelans in RAIS)

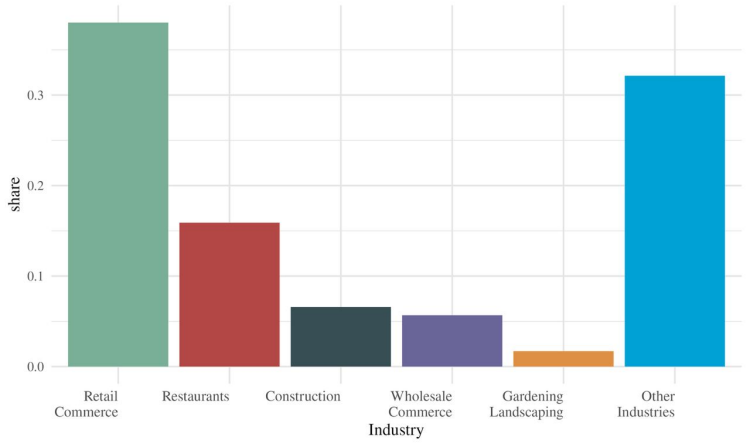
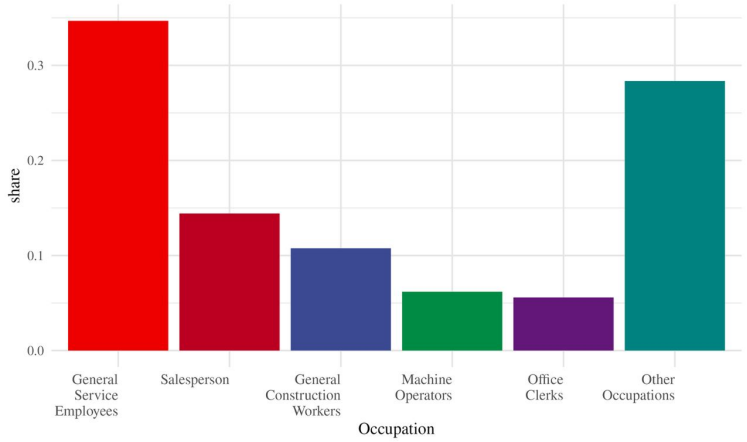


Figure B.7: Occupations of Venezuelans in 2018 (as a percentage of total Venezuelans in RAIS)



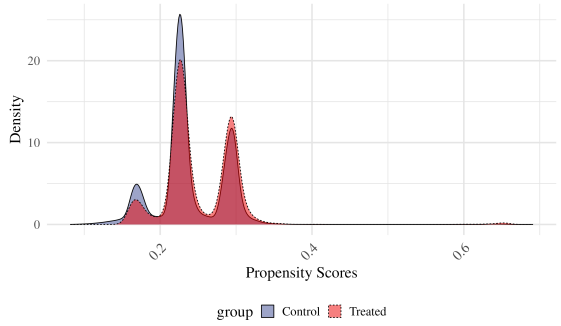


Figure B.8: Covariate Balance for Propensity Score Analysis for College Graduate RAIS Sample

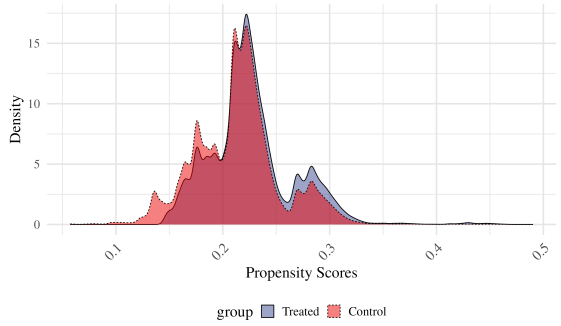


Figure B.9: Covariate Balance for Propensity Score Analysis

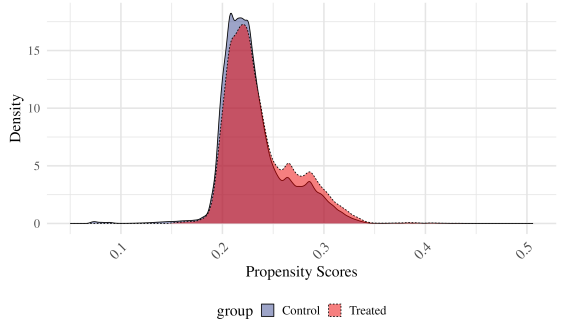


Figure B.10: Covariate Balance for Propensity Score Analysis for High School Graduate RAIS Sample

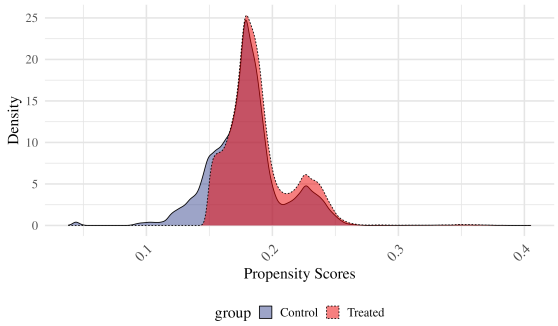


Figure B.11: Covariate Balance for Propensity Score Analysis for RAIS Sample with Less Than High School Education

Appendix C

Additional Content for Chapter 3

C.1 Cluster Choice Analysis

A common drawback of clustering methods is the optimal number of clusters. To increase the robustness of my analysis, I employ the gap statistic method, a widely used technique in cluster analysis (Tibshirani, Walther, and Hastie, 2001), to explore the within variance of clusters and therefore, choose the right number.

The gap statistic compares the total within-cluster variation for different values of k with their expected values under a null reference distribution of the data. I calculate it through the following steps:

1. For each number of clusters K, I compute the within-cluster sum of squares $W_k = \sum_{k=1}^{K} \sum_{i \in k} (w_i - \bar{w}_k)^2$ where k is given cluster, w_i is the observed worker i log-weekly wage, $\bar{w}_k$ the empirical mean of cluster k (centroid).

2. I generate B reference datasets by sampling uniformly from the range of my observed data. For each reference dataset, I compute W_{kb}, the within-cluster sum of squares when clustering the reference data into k clusters.

3. I then compute the gap statistic as:

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \log(W_{kb}) - \log(W_k)$$

4. I use the bootstrapped standard deviation as the standard error across the B reference datasets. I implement this procedure with $B = 500$ to ensure stable estimates.

5. Finally, I choose the optimal number of clusters as the smallest k such that:

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

where s_{k+1} is the standard deviation for $k+1$ clusters.

Figure C.1.1 presents the gap statistic values for different numbers of clusters, ranging from 4 to 25. The green dashed line at 10 clusters represents what my analysis suggests as the optimal number of clusters based on the gap statistic. Due to the large number of observations, standard errors were negligible, barely noticeable in the plot. Nevertheless, there is a notable elbow in the curve, indicating diminishing returns to increasing the number of clusters beyond the value of 10, implying that the benefit of additional clusters in explaining data variability becomes less substantial.

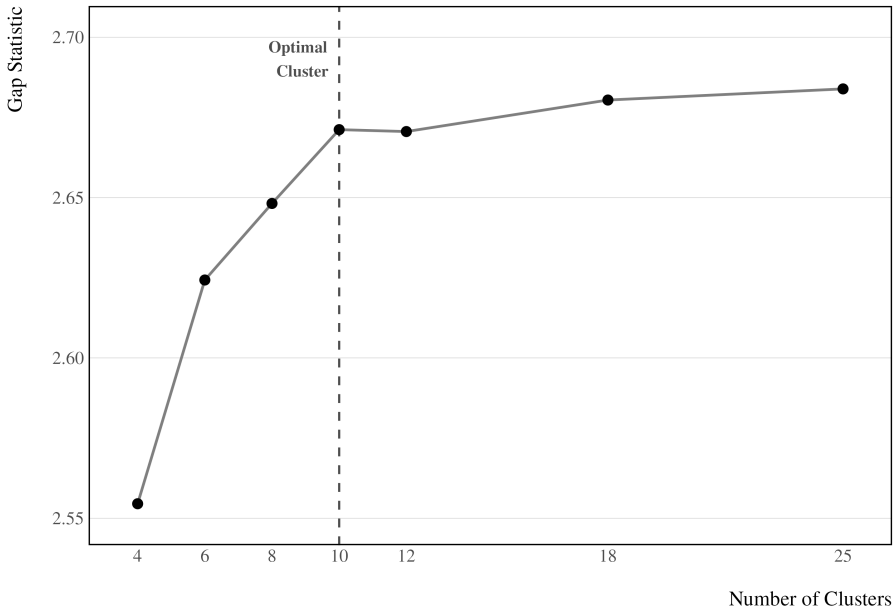


Figure C.1.1: Point Estimate Gap Statistic by Number of Clusters

Note: [1] Vertical line indicates the number of clusters of choice (10). [2] Firm classes estimated by a k-means clustering algorithm using as measurement their empirical cumulative distribution function supported by the ventiles of the population, over six biennials (from 2010-12 to 2015-17). [3] Bootstrapped standard errors were negligible due to the number of observations, with $p < 0.01$.

C.2 Worker Mobility in Firm Clusters

In this subsection, I explore the exogenous mobility assumption of the Gaussian mixture model. This assumption states that the movement of workers should be related to worker types and firm classes, but not directly on earnings. Therefore, the expected wage on unobservables should be zero for job movers.

To test this assumption, first I observe job movers within and across clusters. Movements are considered as long as for each gender workers are changing firms from the first period to the second period. I separate these movements into three categories. Upward movements represent when the worker moves from a lower cluster to an upper cluster. Downward is otherwise. Lateral movements are within cluster job changes.

Figure C.2.1 is constructed by first running a regression following Equation C.12. Than I plot the difference in residuals for every transfer, gender, and data sample cell, discriminated by the movement type. Each dot represents a transfer cell observed in the labor market. The size of the dot indicates how common this particular transfer is.

The figure serves two purposes. First, it addresses obvious trends in job changes, which could indicate that unobservable factors not captured in my model are influencing mobility decisions. A lack of symmetry in the figure would suggest that certain labor market transitions are driven by such unobservables. Second, and equally important, I differentiate these movements by gender to identify any discrepancies that may be endogenous to my model but related to gender differences.

The symmetry plot provides robust evidence supporting the exogenous mobility assumption, which is fundamental to the proper identification of my model. It shows that every movement type possess examples of positive and negative difference in residuals, strongly strongly indicating that job transitions in the labor market are primarily governed by stable firm wage policies and worker characteristics, rather than by time-varying, unobserved factors correlated with wages. Last but not least, the analysis also reveals no apparent gender-specific patterns that would undermine the exogenous mobility assumption for male or female worker samples separately.

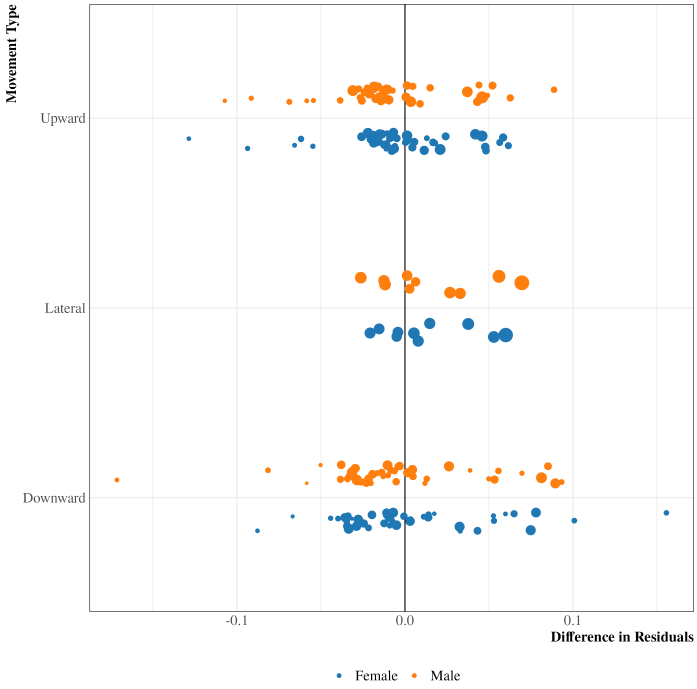


Figure C.2.1: Symmetry plot of job movers' difference in residuals from first to second period.

Note: [1]Dots represent across and within cluster-gender-sample cell movements. [2]Dot size represents each cell's number of observed movements.

C.3 AKM and the Limited Mobility Bias

Here I briefly explain the presence of bias in the AKM estimator of Abowd, Kramarz, and Margolis (1999).

C.3.1 The AKM Model

The AKM is formally written as:

$$w_{it} = X'_{it}\beta + \alpha_i + \phi_{J(i,t)} + \varepsilon_{it} \tag{C.1}$$

where w_{it} are the log earnings of worker i in time t, $X'_{it}\beta$ are exogenous covariates such as age or time period, α_i is the unobserved worker heterogeneity, $J(i,t)$ is an assignment function representing the firm where i works at t, meaning $\phi_{J(i,t)}$ represents the unobserved firm heterogeneity, and ε_{it} is the error term.

Following Bonhomme, Holzheu, et al. (2023), assume N is the number of workers, J the number of firms. For convenience, assume $T = 2$ is the number of time periods. The following assumption must hold:

$$\mathbb{E}[\varepsilon_{it}|X_{11}, \ldots, X_{NT}, j(1,1), \ldots, j(N,T), \alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_J] = 0 \tag{C.2}$$

It is possible, without loss of generality, to rewrite Equation C.1 partialing out $X\beta$ and in vector form. Still following Bonhomme, Holzheu, et al. (2023), I have:

$$W = A\gamma + \varepsilon \tag{C.3}$$

without loss of generality, assume W is subtracted from $X\beta$ and A represents the column-space of worker and firm identifiers.

C.3.2 Connected Set

In matched employer-employee data, the matrix AA' is typically singular, necessitating an additional data cleaning step to ensure a sample of workers and firms that renders AA' non-singular. This step is crucial for the identification of firm and worker effects in additive wage models. For instance, Card, Cardoso, and Kline (2016) estimates gender-specific firm wage effects by isolating the largest dual connected set from the main sample.

This concept of connectivity in the labor market is illustrated in Figure C.3.1, which provides a simplified representation of worker movements across

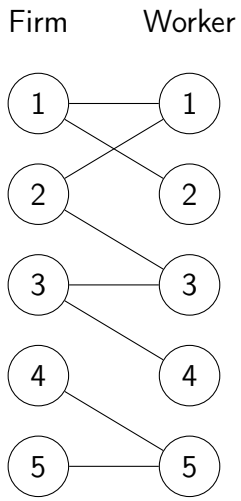


Figure C.3.1: Firm-worker pairs. Firms 1, 2, and 3 are in the largest connected set through workers 1, 2, 3, and 4. Firms 4 and 5 are connected through worker 5 but disjoint from the rest.

[18] For a comprehensive explanation of connected sets and their extraction from data, see Abowd, Creecy, and Kramarz (2002).

firms[18]. It depicts a labor market with five firms and five workers over two time periods. Worker 1 moves from Firm 1 to Firm 2, thereby connecting these two firms. Firm 2 is further connected to Firm 3 through the movement of Worker 3. Firms 4 and 5, while isolated from the first three firms, are connected to each other through Worker 5.

In studies employing matched employer-employee data under additive separability models relying on firm and worker identifiers, researchers typically sample the largest connected set of firms. However, when investigating worker heterogeneity between genders, it is necessary to use the dual connected set, defined as the intersection of the largest connected sets for male and female samples. This approach ensures that firm effects are identified and comparable between both gender groups.

C.3.3 Limited Mobility Bias

The limited mobility bias is a significant concern in the estimation of firm effects, arising from the relative scarcity of job movers in the labor market (Andrews, et al., 2008; Bonhomme, Holzheu, et al., 2023). While this bias does not directly appear in the firm effects estimates from Equation C.1, it manifests in the variance analyses that are commonly employed in the literature to decompose wage inequality.

The sample variances or covariances of interest can be expressed in matrix notation as:

$$\sigma^2 = \gamma' Q \gamma \tag{C.4}$$

where Q is a matrix that depends on the design matrix A.

Andrews, et al. (2008) demonstrated the existence of this bias by decomposing the estimator $\hat{\sigma}^2$:

$$\mathbb{E}[\hat{\sigma}^2|A] = \gamma' Q \gamma + \text{trace}(A(A'A)^{-1}Q(A'A)^{-1}A'\mathbb{V}[\varepsilon|A]) = \sigma^2 + \xi \tag{C.5}$$

where ξ represents the bias term.

Directly correcting for this bias is computationally challenging, as it requires inverting a large matrix, often of dimensions in the hundreds of thousands for firms and millions for workers in typical matched employer-employee datasets. Bonhomme, Holzheu, et al. (2023) have shown that common approximations used in the literature may be insufficient, particularly when relying on fixed effects derived from identifiers. This insufficiency stems from the fact that these approximations often fail to fully account for the complex network structure of worker mobility across firms.

To address these challenges, Bonhomme, Lamadon, and Manresa (2019) proposed a dimension reduction framework. This approach groups firms and workers into a smaller number of classes, thereby increasing the relative probability of observed job changes between groups. While this method effectively mitigates the limited mobility bias, it comes at the cost of imposing additional structure on the estimation model.

C.3.4 Firm Size and Connectivity

Not only does the largest connected set requirement impose a bias due to the rarity of mobility, but it also alters the overall wage distribution of the data. This alteration stems from the fact that the largest connected set tends to include larger firms more frequently than smaller ones. If larger firms differ significantly in their payment schedules and behavior compared to their smaller counterparts, the results derived from such analyses may have limited external validity.

In this section, I provide a formal proof that larger firms are more likely to be included in a connected set of a matched employer-employee dataset. I begin by defining the probability of worker mobility between firms and then demonstrate how this probability scales with firm size.

Without loss of generality, assume $T = 2$. Let $\mathcal{J} = \{1, \ldots, J\}$ be the set of all firms in the economy, and let N_j denote the number of workers in firm j. Define $p_{jj'}$ as the probability that a given worker moves from firm j to another

firm j'. For simplicity, assume that this probability is the same across all workers in the labor market.

Definition 1 (Connected Set) *A connected set $\mathcal{C} \subseteq \mathcal{F}$ is a subset of firms such that for any two firms $j, j' \in \mathcal{C}$, there exists a sequence of firms $j_1, \ldots, j_C \in \mathcal{C}$ with $j_1 = j$, $j_C = j'$ and for each connection $c \in \{1, \ldots, C-1\}$, there is at least one worker who has been employed in both i_c and i_{c+1}.*

Lemma 1 (Probability of Observed Mobility) *The probability of observing at least one worker moving from firm j to firm j' is:*

$$P(j \to j') = 1 - (1 - p_{jj'})^{N_j} \tag{C.6}$$

Proof. The probability of a single worker not moving from j to j' is $(1 - p_{jj'})$. Assume that for all N_j workers not to move, this must occur independently for each worker. Thus, the probability that no workers are moving is $(1 - p_{jj'})^{N_j}$, and the probability that at least one worker is moving is the complement of this event. ■

Theorem 1 *The probability of a firm being part of the connected set is increasing in firm size.*

Proof. For firm j to be part of the connected set, it must have at least one worker moving to or from another firm in the set. The probability of firm j being connected is:

$$P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j} (1 - P(j \to j')) \cdot \prod_{j'' \neq j} (1 - P(j'' \to j)) \tag{C.7}$$

Substituting the result from Lemma 1:

$$P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j} (1 - p_{jj'})^{N_j} \cdot \prod_{j'' \neq j} (1 - p_{j''j})^{N_{j''}} \tag{C.8}$$

To show that this probability increases with firm size, take the derivative with respect to N_j:

Taking the derivative with respect to N_j:

$$\frac{\partial P(j \in \mathcal{C})}{\partial N_j} = -\left(\prod_{j' \neq j} (1 - p_{jj'})^{N_j} \cdot \prod_{j'' \neq j} (1 - p_{j''j})^{N_{j''}} \right) \cdot \sum_{j' \neq j} \log(1 - p_{jj'}) \tag{C.9}$$

Since $0 < p_{jj'} < 1$, we have $\log(1 - p_{jj'}) < 0$, and thus $\frac{\partial P(j \in \mathcal{C})}{\partial N_j} > 0$. ■

Corollary 1 *As firm size approaches infinity, the probability of being in the connected set approaches 1:*

$$\lim_{N_j \to \infty} P(j \in \mathcal{C}) = 1 \tag{C.10}$$

Proof. As $N_j \to \infty$, $(1 - p_{jj'})^{N_j} \to 0$ since $0 < p_{jj'} < 1$. Therefore, the product $\prod_{j \neq j'} (1 - p_{jj'})^{N_j} \to 0$, and consequently, $P(j \in \mathcal{C}) \to 1$. ■

Other firms approaching infinity Another consequence of this proof is when the size of any other firm N_{j_0} approaches infinity while N_j remains finite, $P(j \in \mathcal{C})$ also approaches 1, but at a slower rate. This is because:

$$\lim_{N_{j_0} \to \infty} P(j \in \mathcal{C}) = 1 - \prod_{j' \neq j} (1 - p_{jj'})^{N_j} \cdot 0 \cdot \prod_{j'' \neq j, j'' \neq j_0} (1 - p_{jj''})^{N_{j''}} = 1 \tag{C.11}$$

However, this convergence is slower than when $N_j \to \infty$ because only one term in the product approaches zero, rather than all terms involving N_j.

Empirical Evidence

The theoretical framework is substantiated by empirical evidence presented in Figures C.3.2 and C.3.3, which illustrate the differences between the full sample and the largest dual connected set (LDCS) in terms of firm size and wage distributions.

Figure C.3.2 reveals a stark contrast in the distribution of workforce size between the full sample and the LDCS. The LDCS exhibits a symmetrical distribution shifted significantly to the right, with a mean firm size of approximately 194 workers, compared to the full sample's mean of 27 workers. This rightward shift is accompanied by increased variability, with the standard deviation in the LDCS being almost four times higher than in the original sample. The median firm size in the LDCS is also notably higher, underscoring the overrepresentation of larger firms in the connected set.

The wage distribution, as depicted in Figure C.3.3, further emphasizes the discrepancies between the full sample and the LDCS. The mean log wage in the full sample is 1.87, with a standard deviation of 0.446, while the LDCS shows a substantially higher mean log wage of 2.32 and a larger standard deviation of 0.656. This upward shift in both moments indicates a clear upward bias in wage levels within the connected set. The increased standard deviation in the LDCS also points to greater wage dispersion among the firms included in this subset.

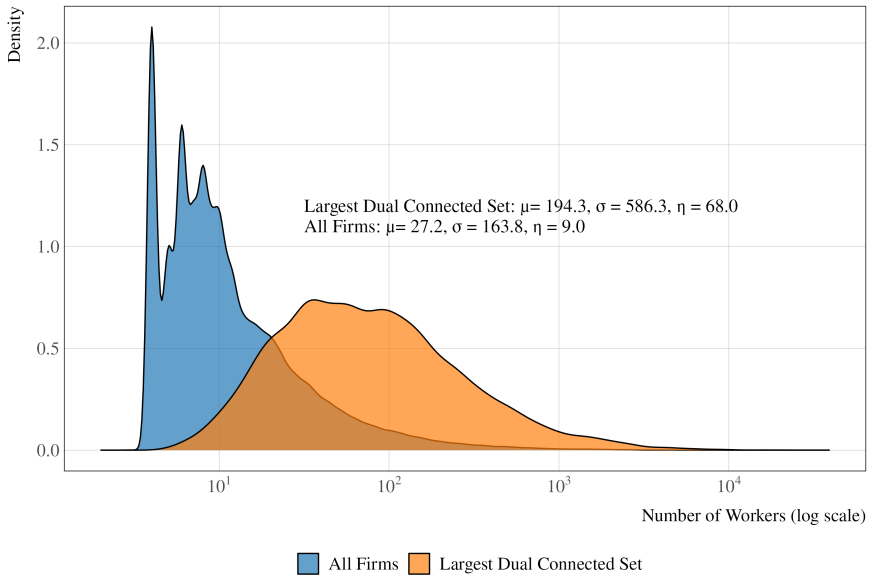


Figure C.3.2: All Firms and The LDCS Number of Workers Distributions

Note: [1]Distributions generated from the six biennial samples, using the full firm set, and the largest dual connected set of firms. [2]μ is the mean of the distribution, σ represents the standard deviation, η represents the median number of workers per firm.

These findings have some implications for the estimation and interpretation of gender wage gaps. LDCS exhibits a larger gender wage gap (δ = -0.317) compared to the entire sample (δ = -0.237), suggesting that studies using the connected set may overestimate the overall wage disparity. This overestimation likely stems from the LDCS that captures wage dynamics primarily in larger, more established firms where gender wage differences might be more pronounced. The exclusion of smaller, potentially lower-paying firms that do not meet the connectivity requirements for AKM-style fixed effects estimation contributes to this bias.

Researchers should exercise caution when generalizing results from the connected set to the broader labor market. The LDCS, while providing the necessary conditions for certain econometric techniques, may not fully represent the wage structures and gender dynamics present in smaller or less connected firms. This limitation is particularly important when studying labor markets with a significant proportion of small enterprises or sectors with limited inter-firm mobility.

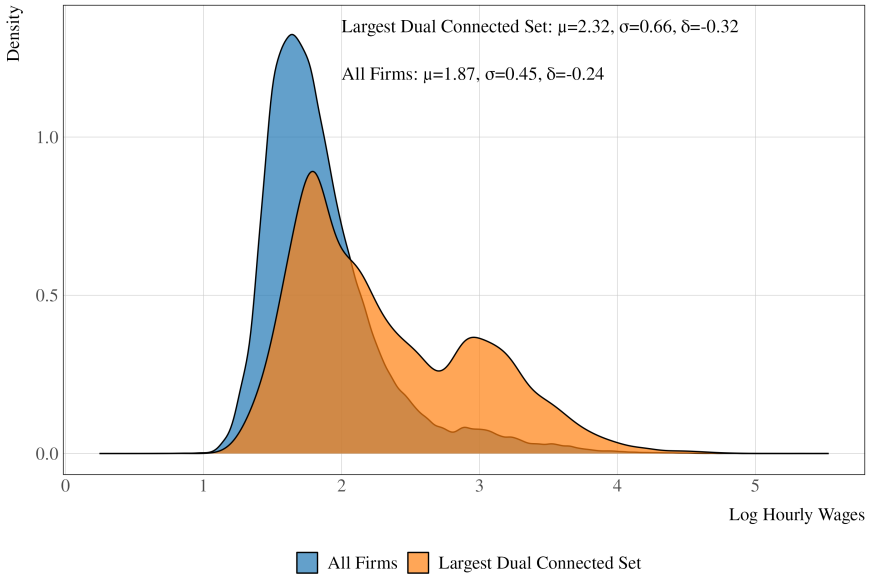


Figure C.3.3: All Firms and The LDCS Log-Weekly Wage Distributions

Note: [1]Distributions generated from the six biennial samples, using the full firm set, and the largest dual connected set of firms. [2]μ is the mean of the distribution, σ represents the standard deviation, δ represents the average female-male wage gap.

C.4 Applying Estimated Clusters in a Linear Framework

Here I provide two exercises to show the mixture model can be used as a plugin estimator of unobserved heterogeneity in a linear regression setting.

First, I employ the estimated firm clusters to estimate firm heterogeneity contribution to the gender wage gap in a classical AKM-KOB from Card, Cardoso, and Kline (2016). The novelty is that I keep social identifiers as the worker plugin estimators, however, I leverage the k-means clusters of firms to avoid data trimming.

Second, I provide a variance decomposition analysis, comparing classical AKM decomposition with the clustered AKM provided previously and the BLM decomposition, where I replace worker social identifiers with their respective mixture membership.

C.4.1 Estimating Firm Contribution to the Gender Wage Gap

Here, I employ the estimated firm clusters in an AKM-KOB classical linear framework to examine their contribution to the gender wage gap. The advantage of this approach is that by grouping firms with similar wage structures, I am not required to filter the dual connected set from the data.

The model I employ has strong similarities to the typical AKM framework. A key distinction lies in the treatment of firm heterogeneity. Instead of utilizing the standard firm assignment function $J(i,t)$, which maps each worker-year observation to a specific firm, I introduce cluster assignment function $K(i,t)$. This function maps worker-year observations to firm clusters, thereby reducing the dimensionality of the firm entities.

Formally, the wage equation can be expressed as:

$$w_{it} = \alpha_i + \psi^g_{K(i,t)} + X^{g\prime}_{it}\beta + \varepsilon_{it} \tag{C.12}$$

where w_{it} is the log wage of worker i in year t, α_i is the worker fixed effect, $\psi_{K(i,t)}$ is the effect of the firm cluster to which worker i's employer belongs in year t, X_{it} is a vector of time-varying covariates, and ε_{it} is the error term. Superscript g indicates that I apply the regression to both the male and female samples.

In particular, the key parameter of interest in this analysis is the difference in firm endowment between male and female workers, derived from the KOB decomposition. This parameter represents the differential distribution of workers across firm clusters by gender, serving as a preliminary measure of gender-

specific assortative matching patterns in the labor market, although focusing solely on firm heterogeneity.

Specifically, I adapt the KOB decomposition to the context of firm cluster effects, formalizing it as:

$$\underbrace{E[\psi^{f}_{K(i,t)}|female] - E[\psi^{m}_{K(i,t)}|male]}_{\substack{\text{Firm Cluster} \\ \text{Contribution to the} \\ \text{Gender Wage Gap}}} = \tag{C.13}$$

$$\text{Bargaining: } \underbrace{\frac{1}{2}\sum_{x \in F,M} (E[\psi^{F}_{K(i,t)} - \psi^{M}_{K(i,t)}|g = x])}_{\substack{\text{Unexplained Portion} \\ \text{(Difference in Returns)}}} \tag{C.14}$$

$$\text{Sorting: } + \underbrace{\frac{1}{2}\sum_{x \in F,M} (E[\psi^{x}_{K(i,t)}|g = F] - E[\psi^{x}_{K(i,t)}|g = M])}_{\substack{\text{Explained Portion} \\ \text{(Difference in Distributions)}}} \tag{C.15}$$

where the left-hand side of the equation represents the total contribution of firm cluster effects to the gender wage gap. This contribution is decomposed into two components: the bargaining effect and the sorting effect.

The bargaining effect, following the terminology of Card, Cardoso, and Kline (2016), captures the portion of the gap attributable to differences in the estimated firm premia between men and women, holding the distribution of firms constant. This effect is computed as the average of two counterfactuals: one using the observed female distribution of firms and another using the observed male distribution.

The sorting effect, conversely, measures the portion of the gap that arises from differences in the distribution of men and women across firm clusters, assuming gender-neutral firm effects. This is the difference in endowments of the Oaxaca decomposition. This effect is also computed as the average of two counterfactuals: one using the estimated male returns to firm clusters and another using the estimated female returns.

By averaging these counterfactuals for each component, as suggested by Casarico and Lattanzio (2024), I obtain robust estimates that account for potential sensitivity to the choice of reference group. Unless otherwise specified, the reported bargaining and sorting effects refer to these averaged estimates.

Normalizing Firm Effects

There is established practice in the AKM literature on gender wage gaps in using the hotel and restaurant industry as a reference[19]. This sector is often chosen due to its typical low wage premia and high turnover rates, suggesting minimal rents (Card, Cardoso, and Kline, 2016; Coudin, Maillard, and Tô, 2018).

[19] Examples are Cruz and Rau (2022) and Casarico and Lattanzio (2024)

Although class 2 firms have a slightly higher proportion of hotels and restaurants (1 percent against 0.6 percent in class 1)[20], I argue that class 1 is the most appropriate reference for several reasons. First, conditional average wages on firm fluster are fairly linear, with class 1 exhibiting the lowest average wage premium in my pooled regression, aligning with the theoretical expectation that the reference group should represent firms offering minimal rents. When employing the AKM regression with class 1 as the reference, resulting fixed effects preserve the linear behavior, with no class exhibiting negative estimates.

[20] See Figure C.5.1 for the estimated premia and the proportion of hotels and restaurants per firm class

Oaxaca Decomposition of Firm Cluster Effects

The main findings derived from the estimation of Equation C.12 are summarized in Table C.4.1. These results represent weighted average firm cluster effects estimates obtained from the six separate biennial samples.

First, considering the overall sample, there is a substantial gender wage gap of 23.7 log points. Firm-specific factors, in Column (2), are estimated around 3.1 log points. As explained by Card, Cardoso, and Kline (2016), this component can be interpreted as the difference in rent payment relative to firm class 1. The component accounts for 13 percent of this total gap.

This contribution can be further decomposed into the sorting component, the difference in male and female worker distributions considering a gender-neutral relative rent. In this case, the total difference is evaluated at 2.1 log points, corresponding to approximately 9 percent of the total gender wage gap. Likewise, the bargaining channel is the average difference in the estimated premium, assuming both genders possess the same firm share. This channel is measured at 1 log point, about 4 percent of the overall gender wage gap.

Following the literature, the lower rows of Table C.4.1 show that the gender wage gap increases dramatically with age, with firms playing a role in this increase, since for individuals older than 50, 23 percent of the 32.6 wage gap is due to estimated firm class effects. Notably, while the sorting component remains relatively stable across age groups (ranging from 0.8 to 2.7 log points), the Bargaining component increases substantially, from 0.7 log points for the youngest group to 4.7 log points for the oldest. This pattern suggests that as workers age, differences in how firms compensate men and women in similar positions become increasingly important in explaining the gender wage gap.

Table C.4.1: Firm Decomposition of the Gender Wage Gap: Overall and by Subgroups

Group	Total Gap	Contribution to Gender Wage Gap		
		Firm Components	Sorting Components	Bargaining Components
	(1)	(2)	(3)	(4)
All	−0.237	−0.033	−0.022	−0.01
		(0.14)	(0.09)	(0.04)
By age group:				
Up to age 30	−0.092	−0.014	−0.009	−0.005
		(0.16)	(0.1)	(0.06)
Ages 31-50	−0.305	−0.041	−0.029	−0.012
		(0.14)	(0.09)	(0.04)
Over age 50	−0.326	−0.076	−0.028	−0.048
		(0.23)	(0.09)	(0.15)
By education group:				
No High school	−0.297	−0.046	−0.035	−0.011
		(0.15)	(0.12)	(0.04)
High school	−0.233	−0.021	−0.025	0.004
		(0.09)	(0.11)	(−0.02)
College	−0.348	−0.075	−0.03	−0.046
		(0.22)	(0.08)	(0.13)

Notes: [1]This table presents the decomposition of the gender wage gap into components attributable to clustered firm-specific factors, using Equation C.12. Column (1) shows the total female-male wage gap in means. Column (2) presents the total contribution of firm-specific factors. Columns (3) and (4) further decompose the firm premium contribution into a sorting (explained) and a bargaining (unexplained) components, respectively. [2]Numbers in parenthesis represent the fraction of the overall gender wage gap that is attributed to the source described in column heading. [3]Results are an weighted average of the six biennial samples.

The analysis by education level reveals a pattern that aligns with findings from the U.S. labor markets, given the wage gap is more prominent among college-educated workers (34.8 log points), albeit the smallest among those with a high school education (23.3 log points). Workers without a high school diploma fall in between, with a gap of 29.7 log points. Interestingly, the contribution of firm-specific factors to the gap follows a similar pattern, being the highest for college-educated workers (5.6 log points) and the lowest for high school graduates (2.8 log points).

For workers without a high school education and those with a high school diploma, the sorting component dominates the bargaining component. This suggests that for these groups, the allocation of women across firms plays a more significant role in the gender wage gap than within-firm differences in compensation. However, the picture changes for college-educated workers. In this group, the bargaining component (3.0 log points) marginally exceeds the sorting component (2.6 log points), indicating that within-firm differences in compensation between men and women become more pronounced as individuals accumulate human capital. Individuals with higher levels of human capital tend to be allocated to more specialized occupations, often within larger firms. However, women may be concentrated at lower paying occupations compared to male counterparts, resulting in workers with greater human capital accumulation securing positions that ultimately translates into heterogeneous bargaining effects within large firms.

Other Cluster Choices and Classical AKM

Table C.4.2 presents the robustness analysis of the model, allowing different cluster choices in the clustered AKM (C-AKM), which is the empirical specification of my study, and, for comparison, the classical AKM model. The results span different levels of firm clustering (K = 4, 6, 8, and 10) in my grouped fixed effects approach, as well as the traditional AKM and the baseline clustering under the largest dual connected set.

Under the clustered AKM approach, trimming the data is not required. However, for traditional AKM settings, it is required to extract the largest dual connected set for correct identification. This is reflected by the larger total gap at 31.7 log points, contrasted with the full data 23.7 log points[21].

[21] I reserve the appendix for a comprehensive analysis of the AKM model and the dual connected set requirement. See Section C.3.

There is a modest but consistent increase in firm components when moving from K = 4 to K = 6, increasing from 2.4 to 3.1 log points. This trend suggests that finer firm classifications capture additional nuances in firm-specific contributions to the gender wage gap.

Bargaining Sensitivity to Cluster Normalization

The bargaining channel, however, is the most sensitive to variations in cluster choice, becoming the almost sole driving force of my robustness analysis, with the sorting component practically stable across results. However, it seems that bargaining estimates stabilize between K = 8 and K = 10, indicating that beyond a certain point, further granularity in firm classification yields diminishing returns in terms of explanatory power. Interestingly, $K = 10$ represents the optimal cluster choice in the gap statistics evaluation[22] (Tibshirani, Walther, and Hastie, 2001).

[22] See Section C.1 for a more rigorous discussion on cluster choice.

The bargaining channel sensitivity could be attributed to the normalization procedure. Slicing firms in the data based on wage distribution similarities may keep the sorting of male and female workers in the labor market almost intact, given that women are more concentrated in low-paying firms. However, it may potentially underestimate rents coming from firms grouped at the lowest class that are, in fact, different enough to be categorized separately in more granulated settings. This is confirmed by the striking different estimate coming from running a C-AKM on the LDCS sample. The largest dual connect set is overrepresented by larger firms, therefore, it is possible that only larger firms with strong positive rent sharing for men were kept in the sample, severely underestimating the male worker's firm component returns.

Therefore, if the researcher desires to commit to the linearity assumption of AKM leveraging from the benefits of clustering firms, the best practice is to employ several number of cluster choices to determine the most appropriate configuration. Ideally, it should minimize within-cluster variance, ensuring that firms within each group are sufficiently homogeneous in their payment behavior, minimizing the normalization cost. Moreover, it should maintain enough heterogeneity between clusters to capture meaningful differences between clusters and provide better economic intuition.

Table C.4.2: Firm Decomposition: Different Model Specifications

Group	Total Gap	Contribution to Gender Wage Gap		
		Firm Components	Sorting Components	Bargaining Components
	(1)	(2)	(3)	(4)
K = 4	-0.237	-0.024	-0.019	-0.005
		(0.10)	(0.08)	(0.02)
K = 6	-0.237	-0.031	-0.021	-0.010
		(0.13)	(0.09)	(0.04)
K = 8	-0.237	-0.034	-0.022	-0.013
		(0.15)	(0.09)	(0.06)
K = 10 (Baseline)	-0.237	-0.033	-0.022	-0.010
		(0.14)	(0.09)	(0.04)
AKM	-0.317	-0.046	-0.026	-0.020
		(0.14)	(0.08)	(0.06)
K = 10 on LDCS	-0.317	-0.025	-0.027	0.002
		(0.07)	(0.08)	(-0.01)

Notes: [1]K represents the number of firm clusters in the grouped fixed effects models. [2]AKM refers to the traditional Abowd, Kramarz, and Margolis (1999) specification under the largest dual connected set sample. [3]Numbers in parenthesis represent the fraction of the overall gender wage gap that is attributed to the source described in column heading. [4]Column (1) shows the total female-male wage gap in means. Column (2) presents the total contribution of firm-specific factors. Columns (3) and (4) further decompose the firm premium contribution into a sorting (explained) and a bargaining (unexplained) components, respectively.

C.5 Additional Tables and Figures

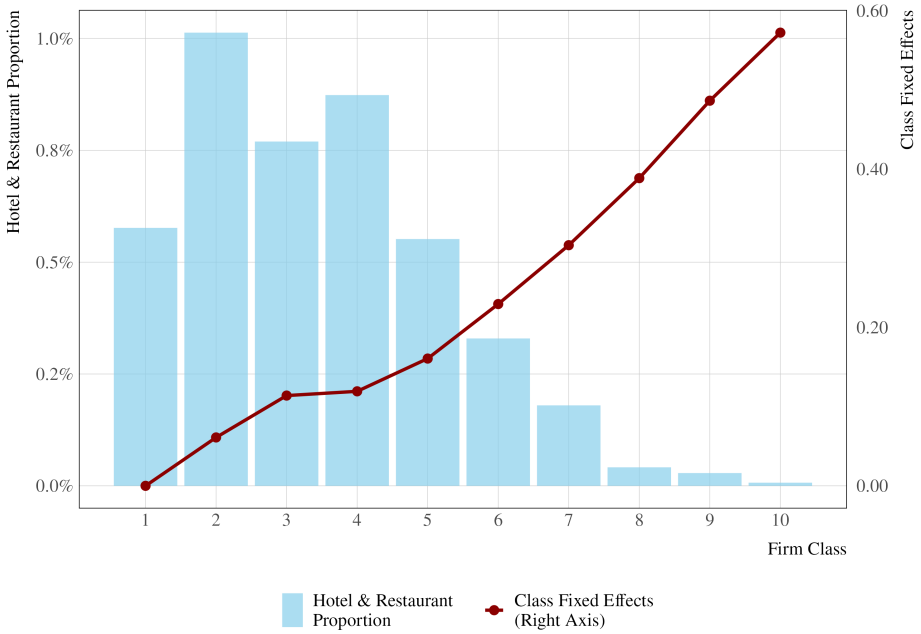


Figure C.5.1: Estimated Effects and Hospitality Industry Proportions Per Firm Class

Note: [1]Firm classes estimated through a kmeans clustering algorithm based on the distribution of logarithmic wages. [2]Hotels and restaurants extracted from code 55 and 56 of the Brazilian *CNAE* code of economic activities.

Note: [1] Descriptive statistics calculated from the first year of each biennial sample's largest dual connected set (2010-2015). [2] Percentages may not sum to 100% due to rounding. [3] The number of firms is the same for each gender since every firm in the cleaned sample employs both male and female workers.

Features	**Female Workers**	**Male Workers**
Firm Characteristics		
Number of Firms	24 500	24 500
Firms with $\geq$ 10 Workers	21 873	21 873
Firms with $\geq$ 50 Workers	12 835	12 835
Mean Firm Size	365	365
Median Firm Size	55	55
Worker Characteristics		
Education (%)		
Dropout	19	22
High School Graduates	46	41
Some College	35	37
Age (%)		
$<$ 30	41	38
31–50	51	50
$\geq$ 51	7	10
Sector of Employment (%)		
Primary	1	2
Manufacturing	19	28
Construction	1	2
Trade	15	15
Services	65	54
Occupation (%)		
Scientific and Liberal Arts	15	16
Technicians	14	14
Administrative	32	19
Managers	5	8
Traders	22	19
Rural	1	2
Factory	18	23
Labor Market Outcomes		
Mean Experience (years)	4.56	5.24
Mean log hourly Wage	2.199	2.516
Variance of log hourly Wage	0.639	0.802
Worker-Year Observations	4 464 653	4 469 690
Number of Workers	1 831 797	1 812 494
Gender Fraction (%)	50	50

Table C.5.2: Descriptive Statistics of Lower Firm Classes

	class 1		class 2		class 3		class 4		class 5	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	74 865	74 865	101 454	101 454	83 715	83 715	79 165	79 165	80 060	80 060
Firms with ≥10 Workers	21 976	21 976	35 851	35 851	32 126	32 126	28 479	28 479	37 413	37 413
Firms with ≥50 Workers	2382	2382	5259	5259	6176	6176	5284	5284	8437	8437
Mean Firm Size	18	18	18	18	22	22	23	23	31	31
Median Firm Size	5	5	6	6	7	7	6	6	9	9
Dropout	0.54	0.52	0.36	0.44	0.28	0.40	0.31	0.40	0.27	0.39
High School Graduates	0.41	0.43	0.57	0.51	0.64	0.54	0.55	0.49	0.60	0.52
Some College	0.05	0.05	0.07	0.05	0.08	0.06	0.14	0.11	0.13	0.09
Age (<30)	0.33	0.40	0.43	0.42	0.45	0.44	0.41	0.39	0.43	0.40
Age 31-50	0.52	0.42	0.46	0.42	0.47	0.42	0.48	0.44	0.48	0.46
Age (≥51)	0.13	0.17	0.09	0.14	0.07	0.12	0.09	0.15	0.08	0.13
Primary Sector	0.05	0.09	0.03	0.05	0.02	0.04	0.03	0.05	0.02	0.03
Manufacturing	0.11	0.11	0.16	0.15	0.17	0.18	0.17	0.20	0.27	0.30
Construction	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Trade	0.15	0.22	0.32	0.35	0.51	0.50	0.25	0.29	0.36	0.36
Services	0.68	0.57	0.49	0.44	0.29	0.27	0.54	0.44	0.35	0.31
Scientific and Liberal Arts	0.02	0.02	0.02	0.02	0.02	0.02	0.05	0.04	0.04	0.03
Technicians	0.03	0.05	0.04	0.05	0.04	0.05	0.06	0.08	0.08	0.07
Administrative	0.20	0.14	0.34	0.15	0.39	0.17	0.36	0.19	0.36	0.16
Managers	0.02	0.03	0.03	0.05	0.04	0.06	0.04	0.05	0.03	0.05
Traders	0.56	0.44	0.39	0.40	0.33	0.36	0.32	0.29	0.27	0.29
Rural	0.04	0.10	0.03	0.05	0.02	0.03	0.02	0.05	0.01	0.02
Factory	0.13	0.22	0.15	0.28	0.16	0.32	0.15	0.30	0.21	0.38
Mean experience (years)	2.789	3.031	2.818	3.136	3.238	3.521	3.405	3.959	3.635	4.033
Mean Log-Wage	1.358	1.466	1.514	1.624	1.660	1.775	1.667	1.860	1.806	1.949
Variance of Log-Wage	0.061	0.101	0.058	0.087	0.074	0.117	0.156	0.206	0.113	0.155
Worker-years observations	775 686	540 010	973 659	888 531	953 073	921 694	942 303	881 295	1 143 801	1 317 477
Number of Workers	437 700	311 953	604 485	544 422	607 113	593 804	626 618	583 517	667 466	757 214
Fraction of Women	0.59	0.41	0.52	0.48	0.51	0.49	0.52	0.48	0.46	0.54

Table C.5.3: Descriptive Statistics of Upper Firm Classes

	class 6		class 7		class 8		class 9		class 10	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	70 196	70 196	46 157	46 157	25 680	25 680	17 077	17 077	7762	7762
Firms with ≥10 Workers	35 405	35 405	24 479	24 479	14 465	14 465	11 627	11 627	5347	5347
Firms with ≥50 Workers	8915	8915	7124	7124	4946	4946	4855	4855	2448	2448
Mean Firm Size	40	40	55	55	81	81	110	110	148	148
Median Firm Size	10	10	10	10	12	12	20	20	20	20
Dropout	0.19	0.31	0.12	0.23	0.07	0.12	0.02	0.05	0.01	0.02
High School Graduates	0.57	0.53	0.52	0.51	0.40	0.43	0.15	0.23	0.08	0.11
Some College	0.24	0.16	0.36	0.26	0.54	0.44	0.83	0.73	0.91	0.87
Age (<30)	0.41	0.37	0.39	0.36	0.39	0.35	0.40	0.33	0.31	0.25
Age 31-50	0.50	0.49	0.52	0.51	0.53	0.53	0.52	0.54	0.58	0.60
Age (≥51)	0.08	0.12	0.08	0.11	0.07	0.10	0.06	0.11	0.09	0.13
Primary Sector	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Manufacturing	0.29	0.36	0.21	0.34	0.14	0.29	0.12	0.19	0.22	0.27
Construction	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Trade	0.23	0.25	0.16	0.21	0.12	0.15	0.08	0.08	0.13	0.13
Services	0.47	0.36	0.61	0.43	0.72	0.53	0.78	0.70	0.63	0.58
Scientific and Liberal Arts	0.09	0.06	0.14	0.10	0.23	0.17	0.32	0.31	0.36	0.36
Technicians	0.13	0.12	0.19	0.15	0.23	0.20	0.13	0.17	0.14	0.17
Administrative	0.34	0.18	0.34	0.19	0.33	0.21	0.36	0.23	0.29	0.19
Managers	0.04	0.05	0.04	0.05	0.06	0.08	0.11	0.14	0.16	0.21
Traders	0.21	0.20	0.16	0.16	0.10	0.11	0.06	0.06	0.04	0.04
Rural	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Factory	0.19	0.39	0.12	0.34	0.06	0.23	0.02	0.09	0.01	0.03
Mean experience (years)	4.264	4.687	4.664	5.089	5.049	5.567	5.367	6.218	5.889	6.418
Mean Log-Wage	1.990	2.162	2.258	2.399	2.575	2.729	2.986	3.201	3.497	3.749
Variance of Log-Wage	0.210	0.253	0.302	0.340	0.377	0.463	0.374	0.470	0.416	0.479
Worker-years observations	1 263 982	1 561 044	1 168 436	1 361 439	990 776	1 080 347	819 258	1 051 482	472 259	680 152
Number of Workers	675 731	823 446	587 173	686 629	449 740	499 661	334 985	427 467	179 633	251 992
Fraction of Women	0.45	0.55	0.46	0.54	0.48	0.52	0.44	0.56	0.41	0.59

Table C.5.4: Descriptive Statistics of Lower Worker Types

	type 1		type 2		type 3		type 4		type 5	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	142 765	115 262	195 800	167 929	244 655	237 647	160 499	173 160	170 689	213 850
Firms with ≥10 Workers	18 551	18 551	37 529	37 529	73 454	73 454	33 732	33 732	54 100	54 100
Firms with ≥50 Workers	2262	2262	5428	5428	12 445	12 445	5354	5354	9744	9744
Mean Firm Size	7	7	9	9	14	14	9	9	13	13
Median Firm Size	2	2	3	3	4	4	3	3	3	3
Dropout	0.44	0.46	0.35	0.41	0.30	0.38	0.28	0.35	0.16	0.33
High School Graduates	0.51	0.49	0.57	0.52	0.61	0.54	0.57	0.51	0.58	0.53
Some College	0.05	0.05	0.08	0.07	0.09	0.07	0.15	0.14	0.26	0.13
Age (<30)	0.40	0.52	0.46	0.55	0.45	0.49	0.41	0.41	0.44	0.39
Age 31-50	0.47	0.33	0.44	0.33	0.46	0.38	0.48	0.44	0.48	0.48
Age (≥51)	0.11	0.13	0.08	0.11	0.08	0.12	0.09	0.13	0.07	0.12
Primary Sector	0.04	0.06	0.02	0.04	0.02	0.03	0.02	0.04	0.01	0.03
Manufacturing	0.17	0.17	0.20	0.20	0.23	0.24	0.20	0.24	0.21	0.27
Construction	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Trade	0.21	0.29	0.29	0.34	0.33	0.34	0.26	0.29	0.26	0.28
Services	0.57	0.47	0.47	0.41	0.42	0.38	0.51	0.42	0.52	0.41
Scientific and Liberal Arts	0.01	0.01	0.02	0.02	0.02	0.02	0.04	0.04	0.07	0.04
Technicians	0.04	0.05	0.04	0.06	0.05	0.06	0.08	0.07	0.14	0.11
Administrative	0.27	0.21	0.34	0.22	0.36	0.19	0.37	0.18	0.42	0.19
Managers	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.05	0.04	0.03
Traders	0.47	0.38	0.38	0.36	0.33	0.34	0.31	0.28	0.18	0.23
Rural	0.04	0.07	0.02	0.04	0.02	0.04	0.02	0.04	0.00	0.02
Factory	0.16	0.26	0.18	0.29	0.20	0.32	0.16	0.34	0.14	0.37
Mean experience (years)	2.552	2.510	2.645	2.530	3.034	2.993	3.517	3.765	4.190	4.230
Mean Log-Wage	1.395	1.433	1.519	1.550	1.624	1.692	1.770	1.955	2.016	2.028
Variance of Log-Wage	0.032	0.037	0.052	0.062	0.061	0.174	0.248	0.463	0.057	0.054
Worker-years observations	742 040	429 828	1 233 122	834 399	2 216 912	1 829 454	951 776	999 212	1 378 256	1 865 438
Number of Workers	634 290	383 994	1 006 451	710 878	1 517 927	1 330 372	808 545	854 727	990 969	1 316 173
Fraction of Women	0.63	0.37	0.60	0.40	0.55	0.45	0.49	0.51	0.42	0.58

Table C.5.5: Descriptive Statistics of Upper Worker Types

	type 6		type 7		type 8		type 9		type 10	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Firms	132 151	167 588	87 574	115 373	51 576	62 286	29 157	39 610	14 021	21 175
Firms with ≥10 Workers	40 967	40 967	32 009	32 009	15 816	15 816	11 785	11 785	6470	6470
Firms with ≥50 Workers	7489	7489	6674	6674	3481	3481	2998	2998	1676	1676
Mean Firm Size	13	13	15	15	15	15	21	21	22	22
Median Firm Size	3	3	3	3	2	2	3	3	3	3
Dropout	0.12	0.26	0.05	0.18	0.03	0.08	0.01	0.04	0.00	0.01
High School Graduates	0.43	0.48	0.28	0.40	0.14	0.25	0.07	0.15	0.04	0.07
Some College	0.45	0.26	0.67	0.42	0.83	0.67	0.92	0.82	0.96	0.92
Age (<30)	0.38	0.32	0.35	0.29	0.28	0.24	0.19	0.17	0.08	0.08
Age 31-50	0.53	0.54	0.56	0.57	0.62	0.61	0.68	0.66	0.73	0.69
Age (≥51)	0.07	0.13	0.07	0.12	0.09	0.13	0.11	0.15	0.16	0.21
Primary Sector	0.01	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Manufacturing	0.16	0.28	0.15	0.29	0.16	0.28	0.18	0.29	0.18	0.26
Construction	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Trade	0.19	0.23	0.14	0.18	0.12	0.14	0.11	0.13	0.12	0.13
Services	0.64	0.45	0.69	0.50	0.71	0.55	0.69	0.56	0.67	0.58
Scientific and Liberal Arts	0.17	0.09	0.31	0.18	0.42	0.31	0.44	0.38	0.38	0.36
Technicians	0.22	0.15	0.22	0.20	0.15	0.19	0.11	0.17	0.07	0.09
Administrative	0.35	0.19	0.29	0.19	0.22	0.15	0.19	0.13	0.15	0.10
Managers	0.06	0.06	0.08	0.08	0.12	0.13	0.20	0.20	0.38	0.41
Traders	0.13	0.16	0.07	0.11	0.08	0.08	0.05	0.05	0.02	0.03
Rural	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Factory	0.07	0.33	0.03	0.24	0.01	0.12	0.01	0.06	0.00	0.02
Mean experience (years)	4.943	5.413	5.684	6.145	6.201	6.661	7.124	7.294	8.331	8.247
Mean Log-Wage	2.347	2.400	2.789	2.765	3.196	3.272	3.624	3.662	4.157	4.226
Variance of Log-Wage	0.200	0.202	0.074	0.080	0.232	0.213	0.080	0.085	0.103	0.131
Worker-years observations	1 058 198	1 467 569	901 644	1 216 454	486 706	666 147	373 527	612 301	161 052	362 669
Number of Workers	724 998	996 335	529 627	728 982	315 684	437 215	216 089	351 364	91 719	200 886
Fraction of Women	0.42	0.58	0.43	0.57	0.42	0.58	0.38	0.62	0.31	0.69

Table C.5.6: Wage Levels for Males and Females under Different Scenarios

	Baseline		Separable Market		Constant Returns		Constant Firm Allocation	
Group	Female (1)	Male (2)	Female (3)	Male (4)	Female (5)	Male (6)	Female (7)	Male (8)
All	2.10	2.33	2.11	2.30	2.16	2.25	2.19	2.22
Education								
No HS	1.62	1.92	1.65	1.92	1.72	1.84	1.77	1.80
HS	1.90	2.14	1.92	2.13	1.97	2.07	2.01	2.04
College	2.92	3.27	2.87	3.17	2.94	3.10	3.00	3.03
Age								
<30	1.97	2.06	1.98	2.06	2.01	2.03	2.01	2.03
31-50	2.18	2.49	2.18	2.44	2.25	2.38	2.30	2.33
50>	2.09	2.42	2.08	2.36	2.15	2.29	2.21	2.24
Firm Size								
Firms <10	1.66	1.78	1.65	1.77	1.70	1.72	1.70	1.73
Firms 10-50	1.77	1.90	1.77	1.90	1.82	1.85	1.83	1.84
Firms 51>	1.78	1.99	1.79	1.96	1.83	1.91	1.86	1.88
Occupations								
Hotel and Restaurants	1.62	1.74	1.62	1.74	1.71	1.67	1.67	1.70
Engineers & Economists	2.91	3.30	2.88	3.23	2.98	3.13	3.03	3.08
Managers	3.03	3.35	2.90	3.12	2.96	3.05	2.99	3.02

Notes: [1]All values represent base wages in log scale. [2]Baseline is observed wages. [3]Separable Market assumes no complementarity in worker-firm interactions. [4]Constant Returns equalizes means and variances of realized worker-firm interactions. [5]Constant Firm Allocation equalizes firm-specific probabilities.

Bibliography

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (Feb. 2023). "When Should You Adjust Standard Errors for Clustering?*". In: *The Quarterly Journal of Economics* 138.1, pp. 1–35. ISSN: 0033-5533. DOI: 10.1093/qje/qjac038. URL: https://doi.org/10.1093/qje/qjac038 (visited on Dec. 4, 2023).

Abadie, A., A. Diamond, and J. Hainmueller (June 2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.ap08746. URL: https://doi.org/10.1198/jasa.2009.ap08746 (visited on May 30, 2023).

Abadie, A., A. Diamond, and J. Hainmueller (2015). "Comparative Politics and the Synthetic Control Method". en. In: *American Journal of Political Science* 59.2, pp. 495–510. ISSN: 1540-5907. DOI: 10.1111/ajps.12116. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12116 (visited on May 30, 2023).

Abadie, A. and J. Gardeazabal (Mar. 2003). "The Economic Costs of Conflict: A Case Study of the Basque Country". en. In: *American Economic Review* 93.1, pp. 113–132. ISSN: 0002-8282. DOI: 10.1257/000282803321455188. URL: https://www.aeaweb.org/articles?id=10.1257/000282803321455188 (visited on May 30, 2023).

Abowd, J. M., R. H. Creecy, and F. Kramarz (Mar. 2002). "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data". en. In: *Longitudinal Employer-Household Dynamics Technical Papers*. Number: 2002-06 Publisher: Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org//p/cen/tpaper/2002-06.html (visited on Dec. 8, 2023).

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). "High Wage Workers and High Wage Firms". en. In: *Econometrica* 67.2, pp. 251–333. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00020. URL: https://

onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00020 (visited on Sept. 20, 2023).

Afridi, F., K. Mahajan, and N. Sangwan (Oct. 2022). "The gendered effects of droughts: Production shocks and labor response in agriculture". en. In: *Labour Economics* 78, p. 102227. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2022.102227. URL: https://www.sciencedirect.com/science/article/pii/S0927537122001178 (visited on June 3, 2023).

Aksu, E., R. Erzan, and M. G. Kırdar (June 2022). "The impact of mass migration of Syrians on the Turkish labor market". In: *Labour Economics* 76, p. 102183. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2022.102183. URL: https://www.sciencedirect.com/science/article/pii/S0927537122000744 (visited on Mar. 15, 2024).

Alix-Garcia, J., S. Walker, A. Bartlett, H. Onder, and A. Sanghi (Jan. 2018). "Do refugee camps help or hurt hosts? The case of Kakuma, Kenya". en. In: *Journal of Development Economics* 130, pp. 66–83. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2017.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0304387817300688 (visited on Oct. 20, 2022).

Andrews, M. J., L. Gill, T. Schank, and R. Upward (June 2008). "High Wage Workers and Low Wage Firms: Negative Assortative Matching or Limited Mobility Bias?" In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 171.3, pp. 673–697. ISSN: 0964-1998. DOI: 10.1111/j.1467-985X.2007.00533.x. URL: https://doi.org/10.1111/j.1467-985X.2007.00533.x (visited on May 17, 2024).

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (Dec. 2021). "Synthetic Difference-in-Differences". en. In: *American Economic Review* 111.12, pp. 4088–4118. ISSN: 0002-8282. DOI: 10.1257/aer.20190159. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20190159 (visited on Mar. 29, 2024).

Azeredo, C. (2019). *IBGE - Censo Agro 2017*. pt-br. URL: tinyurl.com/3hd3zpb7 (visited on Dec. 6, 2022).

Azkarate-Askasua, M. and M. Zerecero (Jan. 2023). *Correcting Small Sample Bias in Linear Models with Many Covariates*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4322300. URL: https://papers.ssrn.com/abstract=4322300 (visited on Feb. 5, 2024).

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

0304-3878. DOI: 10.1016/j.jdeveco.2021.102652. URL: https://www.sciencedirect.com/science/article/pii/S0304387821000316 (visited on Oct. 20, 2022).

Barth, E. and H. Dale-Olsen (2003). "Assortative matching in the labour market? Stylised facts about workers and plants". en. In: *Institute for Social Research, Oslo, Norway*.

Becker, G. S. (July 1973). "A Theory of Marriage: Part I". In: *Journal of Political Economy* 81.4. Publisher: The University of Chicago Press, pp. 813–846. ISSN: 0022-3808. DOI: 10.1086/260084. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/260084 (visited on Feb. 11, 2024).

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (Jan. 2019). "Breaking the Glass Ceiling? The Effect of Board Quotas on Female Labour Market Outcomes in Norway". In: *The Review of Economic Studies* 86.1, pp. 191–239. ISSN: 0034-6527. DOI: 10.1093/restud/rdy032. URL: https://doi.org/10.1093/restud/rdy032 (visited on July 30, 2024).

Bertrand, M., C. Goldin, and L. F. Katz (July 2010). "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors". en. In: *American Economic Journal: Applied Economics* 2.3, pp. 228–255. ISSN: 1945-7782. DOI: 10.1257/app.2.3.228. URL: https://www.aeaweb.org/articles?id=10.1257/app.2.3.228 (visited on July 30, 2024).

Black, D. A., A. M. Haviland, S. G. Sanders, and L. J. Taylor (July 2008). "Gender Wage Disparities among the Highly Educated". en. In: *Journal of Human Resources* 43.3. Publisher: University of Wisconsin Press Section: Articles, pp. 630–659. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.43.3.630. URL: https://jhr.uwpress.org/content/43/3/630 (visited on Sept. 12, 2024).

Blau, F. D. and L. M. Kahn (2008). "Women's Work and Wages". en. In: *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK, pp. 1–14. ISBN: 978-1-349-95121-5. DOI: 10.1057/978-1-349-95121-5_2207-1. URL: https://doi.org/10.1057/978-1-349-95121-5_2207-1 (visited on Sept. 12, 2024).

Blinder, A. S. (1973). "Wage Discrimination: Reduced Form and Structural Estimates". In: *The Journal of Human Resources* 8.4. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System], pp. 436–455. ISSN: 0022-166X. DOI: 10.2307/144855. URL: https://www.jstor.org/stable/144855 (visited on Aug. 22, 2023).

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (Apr. 2023). "How Much Should We Trust Estimates of Firm Effects and Worker Sorting?" In: *Journal of Labor Economics* 41.2. Publisher: The University of Chicago Press, pp. 291–322. ISSN: 0734-306X. DOI: 10.1086/720009. URL: https://www.journals.uchicago.edu/doi/full/10.1086/720009 (visited on Mar. 7, 2024).

Bonhomme, S., T. Lamadon, and E. Manresa (2019). "A Distributional Framework for Matched Employer Employee Data". en. In: *Econometrica* 87.3, pp. 699–739. ISSN: 1468-0262. DOI: 10.3982/ECTA15722. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15722 (visited on Aug. 21, 2023).

Bonhomme, S. and E. Manresa (2015). "Grouped Patterns of Heterogeneity in Panel Data". en. In: *Econometrica* 83.3, pp. 1147–1184. ISSN: 1468-0262. DOI: 10.3982/ECTA11319. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA11319 (visited on Sept. 25, 2023).

Borjas, G. J. (Nov. 2003). "The Labor Demand Curve is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market*". In: *The Quarterly Journal of Economics* 118.4, pp. 1335–1374. ISSN: 0033-5533. DOI: 10.1162/003355303322552810. URL: https://doi.org/10.1162/003355303322552810 (visited on Oct. 30, 2022).

Borjas, G. J. (Mar. 2006). "Native Internal Migration and the Labor Market Impact of Immigration". en. In: *Journal of Human Resources* XLI.2. Publisher: University of Wisconsin Press, pp. 221–258. ISSN: 0022-166X, 1548-8004. DOI: 10.3368/jhr.XLI.2.221. URL: http://jhr.uwpress.org/content/XLI/2/221 (visited on Oct. 31, 2022).

Borjas, G. J. (Oct. 2017). "The Wage Impact of the Marielitos: A Reappraisal". en. In: *ILR Review* 70.5. Publisher: SAGE Publications Inc, pp. 1077–1110. ISSN: 0019-7939. DOI: 10.1177/0019793917692945. URL: https://doi.org/10.1177/0019793917692945 (visited on Oct. 15, 2022).

Borjas, G. J., R. B. Freeman, L. F. Katz, J. DiNardo, and J. M. Abowd (1997). "How Much Do Immigration and Trade Affect Labor Market Outcomes?" In: *Brookings Papers on Economic Activity* 1997.1. Publisher: Brookings Institution Press, pp. 1–90. ISSN: 0007-2303. DOI: 10.2307/2534701. URL: https://www.jstor.org/stable/2534701 (visited on Oct. 31, 2022).

Boustan, L. P., M. E. Kahn, P. W. Rhode, and M. L. Yanguas (July 2020). "The effect of natural disasters on economic activity in US counties: A century of data". en. In: *Journal of Urban Economics* 118, p. 103257. ISSN: 0094-1190. DOI: 10.1016/j.jue.2020.103257. URL: https://www.science

direct.com/science/article/pii/S0094119020300280 (visited on June 3, 2023).

Bruns, B. (Apr. 2019). "Changes in Workplace Heterogeneity and How They Widen the Gender Wage Gap". en. In: *American Economic Journal: Applied Economics* 11.2, pp. 74–113. ISSN: 1945-7782. DOI: 10.1257/app.20160664. URL: https://www.aeaweb.org/articles?id=10.1257/app.20160664 (visited on May 29, 2024).

Calderón-Mejía, V. and A. M. Ibáñez (May 2016). "Labour market effects of migration-related supply shocks: evidence from internal refugees in Colombia". In: *Journal of Economic Geography* 16.3, pp. 695–713. ISSN: 1468-2702. DOI: 10.1093/jeg/lbv030. URL: https://doi.org/10.1093/jeg/lbv030 (visited on Mar. 29, 2024).

Callaway, B. and P. H. C. Sant'Anna (Dec. 2021a). "Difference-in-Differences with multiple time periods". en. In: *Journal of Econometrics*. Themed Issue: Treatment Effect 1 225.2, pp. 200–230. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2020.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0304407620303948 (visited on May 23, 2023).

Callaway, B. and P. H. Sant'Anna (Dec. 2021b). "Difference-in-Differences with multiple time periods". en. In: *Journal of Econometrics* 225.2, pp. 200–230. ISSN: 03044076. DOI: 10.1016/j.jeconom.2020.12.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407620303948 (visited on May 23, 2023).

Card, D. (Jan. 1990). "The Impact of the Mariel Boatlift on the Miami Labor Market". en. In: *ILR Review* 43.2. Publisher: SAGE Publications Inc, pp. 245–257. ISSN: 0019-7939. DOI: 10.1177/001979399004300205. URL: https://doi.org/10.1177/001979399004300205 (visited on Oct. 15, 2022).

Card, D. (Nov. 2005). "Is the New Immigration Really so Bad?" In: *The Economic Journal* 115.507, F300–F323. ISSN: 0013-0133. DOI: 10.1111/j.1468-0297.2005.01037.x. URL: https://doi.org/10.1111/j.1468-0297.2005.01037.x (visited on Oct. 30, 2022).

Card, D. (May 2009). "Immigration and Inequality". en. In: *American Economic Review* 99.2, pp. 1–21. ISSN: 0002-8282. DOI: 10.1257/aer.99.2.1. URL: https://www.aeaweb.org/articles?id=10.1257/aer.99.2.1 (visited on Oct. 30, 2022).

Card, D., A. R. Cardoso, and P. Kline (May 2016). "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *". In: *The Quarterly Journal of Economics* 131.2, pp. 633–

686. ISSN: 0033-5533. DOI: 10.1093/qje/qjv038. URL: https://doi.org/10.1093/qje/qjv038 (visited on Dec. 5, 2023).

Card, D., J. Heining, and P. Kline (Aug. 2013). "Workplace Heterogeneity and the Rise of West German Wage Inequality*". In: *The Quarterly Journal of Economics* 128.3, pp. 967–1015. ISSN: 0033-5533. DOI: 10.1093/qje/qjt006. URL: https://doi.org/10.1093/qje/qjt006 (visited on Sept. 25, 2023).

Card, D., I. Schmutte, and L. Vilhuber (Apr. 2023). "Introduction to the Special Issue: Models of linked employer–employee data: Twenty years after "High Wage Workers and High Wage Firms"". en. In: *Journal of Econometrics* 233.2, pp. 333–339. ISSN: 03044076. DOI: 10.1016/j.jeconom.2023.01.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407623000337 (visited on Aug. 5, 2024).

Caruso, G., C. G. Canon, and V. Mueller (Apr. 2021). "Spillover effects of the Venezuelan crisis: migration impacts in Colombia". In: *Oxford Economic Papers* 73.2, pp. 771–795. ISSN: 0030-7653. DOI: 10.1093/oep/gpz072. URL: https://doi.org/10.1093/oep/gpz072 (visited on Oct. 31, 2022).

Casarico, A. and S. Lattanzio (Apr. 2024). "What Firms Do: Gender Inequality in Linked Employer-Employee Data". In: *Journal of Labor Economics* 42.2. Publisher: The University of Chicago Press, pp. 325–355. ISSN: 0734-306X. DOI: 10.1086/723177. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/723177 (visited on Apr. 8, 2024).

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams (Dec. 2014). "Women in Academic Science: A Changing Landscape". en. In: *Psychological Science in the Public Interest* 15.3. Publisher: SAGE Publications Inc, pp. 75–141. ISSN: 1529-1006. DOI: 10.1177/1529100614541236. URL: https://doi.org/10.1177/1529100614541236 (visited on Sept. 12, 2024).

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (May 2017). "Double/Debiased/Neyman Machine Learning of Treatment Effects". en. In: *American Economic Review* 107.5, pp. 261–265. ISSN: 0002-8282. DOI: 10.1257/aer.p20171038. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20171038 (visited on May 23, 2023).

Clemens, M. A. and J. Hunt (Aug. 2019). "The Labor Market Effects of Refugee Waves: Reconciling Conflicting Results". en. In: *ILR Review* 72.4. Publisher: SAGE Publications Inc, pp. 818–857. ISSN: 0019-7939. DOI: 10.1177/0019793918824597. URL: https://doi.org/10.1177/0019793918824597 (visited on Oct. 15, 2022).

Coudin, E., S. Maillard, and M. Tô (2018). *Family, firms and the gender wage gap in France*. eng. Working Paper W18/01. IFS Working Papers. DOI: 10.1920/wp.ifs.2018.W1801. URL: https://www.econstor.eu/handle/10419/200290 (visited on July 7, 2024).

Cruz, G. and T. Rau (Apr. 2022). "The effects of equal pay laws on firm pay premiums: Evidence from Chile". In: *Labour Economics* 75, p. 102135. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2022.102135. URL: https://www.sciencedirect.com/science/article/pii/S0927537122000288 (visited on July 7, 2024).

Delgado-Prieto, L. (n.d.). "Dynamics of Local Wages and Employment: Evidence from the Venezuelan Immigration in Colombia". en. In: ().

Dustmann, C., F. Fasani, T. Frattini, L. Minale, and U. Schönberg (July 2017). "On the economics and politics of refugee migration". In: *Economic Policy* 32.91, pp. 497–550. ISSN: 0266-4658. DOI: 10.1093/epolic/eix008. URL: https://doi.org/10.1093/epolic/eix008 (visited on Mar. 19, 2024).

Dustmann, C., T. Frattini, and I. P. Preston (Jan. 2013). "The Effect of Immigration along the Distribution of Wages". In: *The Review of Economic Studies* 80.1, pp. 145–173. ISSN: 0034-6527. DOI: 10.1093/restud/rds019. URL: https://doi.org/10.1093/restud/rds019 (visited on Oct. 31, 2022).

Echeverría-Estrada Carlos, D. C.-G. (Aug. 2020). *Venezuelan Migrants and Refugees in Latin America and the Caribbean: A Regional Profile*. en. URL: https://www.migrationpolicy.org/research/venezuelans-latin-america-caribbean-regional-profile (visited on Mar. 19, 2024).

Eeckhout, J. and P. Kircher (2018). "Assortative Matching With Large Firms". en. In: *Econometrica* 86.1, pp. 85–132. ISSN: 1468-0262. DOI: 10.3982/ECTA14450. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14450 (visited on Feb. 11, 2024).

Efobi, U. (Aug. 2022). "The Long-Term Labor Market Effect of Drought Exposure: Evidence from Nigeria". In: *The Journal of Development Studies* 58.8, pp. 1531–1549. ISSN: 0022-0388. DOI: 10.1080/00220388.2022.2055464. URL: https://doi.org/10.1080/00220388.2022.2055464 (visited on June 3, 2023).

EIA (2019). *International Analysis: Venezuela*. URL: https://www.eia.gov/international/overview/country/VEN (visited on Nov. 1, 2022).

Fallah, B., C. Krafft, and J. Wahba (June 2019). "The impact of refugees on employment and wages in Jordan". en. In: *Journal of Development Economics* 139, pp. 203–216. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2019.03.009. URL: https://www.sciencedirect.com/science/article/pii/S0304387818310344 (visited on Oct. 20, 2022).

Fernandes, G. W., F. F. Goulart, B. D. Ranieri, M. S. Coelho, K. Dales, N. Boesche, M. Bustamante, F. A. Carvalho, D. C. Carvalho, R. Dirzo, S. Fernandes, P. M. Galetti, V. E. G. Millan, C. Mielke, J. L. Ramirez, A. Neves, C. Rogass, S. P. Ribeiro, A. Scariot, and B. Soares-Filho (July 2016). "Deep into the mud: ecological and socio-economic impacts of the dam breach in Mariana, Brazil". en. In: *Natureza & Conservação* 14.2, pp. 35–45. ISSN: 1679-0073. DOI: 10.1016/j.ncon.2016.10.003. URL: https://www.sciencedirect.com/science/article/pii/S1679007316301104 (visited on Feb. 17, 2023).

Foged, M. and G. Peri (Apr. 2016). "Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data". en. In: *American Economic Journal: Applied Economics* 8.2, pp. 1–34. ISSN: 1945-7782, 1945-7790. DOI: 10.1257/app.20150114. URL: https://pubs.aeaweb.org/doi/10.1257/app.20150114 (visited on May 23, 2023).

Frame, D. E. (July 1998). "Housing, Natural Hazards, and Insurance". en. In: *Journal of Urban Economics* 44.1, pp. 93–109. ISSN: 0094-1190. DOI: 10.1006/juec.1997.2061. URL: https://www.sciencedirect.com/science/article/pii/S0094119097920611 (visited on June 3, 2023).

Friedberg, R. M. (Nov. 2001). "The Impact of Mass Migration on the Israeli Labor Market*". In: *The Quarterly Journal of Economics* 116.4, pp. 1373–1408. ISSN: 0033-5533. DOI: 10.1162/003355301753265606. URL: https://doi.org/10.1162/003355301753265606 (visited on Mar. 19, 2024).

Gabriel, F. Â., R. A. Hauser-Davis, L. Soares, A. C. A. Mazzuco, R. C. C. Rocha, T. D. S. Pierre, E. Saggioro, F. V. Correia, T. O. Ferreira, and A. F. Bernardino (Oct. 2020). "Contamination and oxidative stress biomarkers in estuarine fish following a mine tailing disaster". en. In: *PeerJ* 8. Publisher: PeerJ Inc., e10266. ISSN: 2167-8359. DOI: 10.7717/peerj.10266. URL: https://peerj.com/articles/10266 (visited on June 4, 2023).

Gallen, Y., R. V. Lesner, and R. Vejlin (Jan. 2019). "The labor market gender gap in Denmark: Sorting out the past 30 years". In: *Labour Economics* 56, pp. 58–67. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2018.11.003.

URL: https://www.sciencedirect.com/science/article/pii/S0927537118301234 (visited on Mar. 1, 2024).

Gaure, S. (2014). "Correlation bias correction in two-way fixed-effects linear regression". en. In: *Stat* 3.1, pp. 379–390. ISSN: 2049-1573. DOI: 10.1002/sta4.68. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.68 (visited on Oct. 23, 2023).

Gerard, F., L. Lagos, E. Severnini, and D. Card (Oct. 2021). "Assortative Matching or Exclusionary Hiring? The Impact of Employment and Pay Policies on Racial Wage Differences in Brazil". en. In: *American Economic Review* 111.10, pp. 3418–3457. ISSN: 0002-8282. DOI: 10.1257/aer.20181596. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20181596 (visited on Mar. 7, 2024).

Glitz, A. (Jan. 2012). "The Labor Market Impact of Immigration: A Quasi-Experiment Exploiting Immigrant Location Rules in Germany". In: *Journal of Labor Economics* 30.1. Publisher: The University of Chicago Press, pp. 175–213. ISSN: 0734-306X. DOI: 10.1086/662143. URL: https://www.journals.uchicago.edu/doi/full/10.1086/662143 (visited on Mar. 19, 2024).

Goldin, C. (Apr. 2014). "A Grand Gender Convergence: Its Last Chapter". en. In: *American Economic Review* 104.4, pp. 1091–1119. ISSN: 0002-8282. DOI: 10.1257/aer.104.4.1091. URL: https://www.aeaweb.org/articles?id=10.1257/aer.104.4.1091 (visited on Sept. 20, 2023).

Goldin, C., L. F. Katz, and I. Kuziemko (Dec. 2006). "The Homecoming of American College Women: The Reversal of the College Gender Gap". en. In: *Journal of Economic Perspectives* 20.4, pp. 133–156. ISSN: 0895-3309. DOI: 10.1257/jep.20.4.133. URL: https://www.aeaweb.org/articles?id=10.1257%2Fjep.20.4.133&fbclid=IwAR2YgNWGj5pMpZpEpG5a97ahef4Dwi8Wk7rUpLqqRGPMqwOGOnJYmGvcLys (visited on Sept. 12, 2024).

Goodman-Bacon, A. (Dec. 2021). "Difference-in-differences with variation in treatment timing". en. In: *Journal of Econometrics*. Themed Issue: Treatment Effect 1 225.2, pp. 254–277. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2021.03.014. URL: https://www.sciencedirect.com/science/article/pii/S0304407621001445 (visited on May 23, 2023).

Graham, B. S., G. W. Imbens, and G. Ridder (2014). "Complementarity and aggregate implications of assortative matching: A nonparametric analysis". en. In: *Quantitative Economics* 5.1, pp. 29–66. ISSN: 1759-7331. DOI: 10.

3982/QE45. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/QE45 (visited on May 14, 2024).

Groeger, A., G. León-Ciliotta, and S. Stillman (Feb. 2024). "Immigration, labor markets and discrimination: Evidence from the Venezuelan Exodus in Perú". In: *World Development* 174, p. 106437. ISSN: 0305-750X. DOI: 10.1016/j.worlddev.2023.106437. URL: https://www.sciencedirect.com/science/article/pii/S0305750X23002553 (visited on Mar. 19, 2024).

Groen, J. A. and A. E. Polivka (May 2008). "The Effect of Hurricane Katrina on the Labor Market Outcomes of Evacuees". en. In: *American Economic Review* 98.2, pp. 43–48. ISSN: 0002-8282. DOI: 10.1257/aer.98.2.43. URL: https://www.aeaweb.org/articles?id=10.1257/aer.98.2.43 (visited on June 3, 2023).

Gruetter, M. and R. Lalive (Apr. 2009). "The importance of firms in wage determination". In: *Labour Economics* 16.2, pp. 149–160. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2008.09.001. URL: https://www.sciencedirect.com/science/article/pii/S0927537108001012 (visited on May 17, 2024).

Haider, W. (Jan. 2020). "Estimates of Total Oil & Gas Reserves in The World, Future of Oil and Gas Companies and SMART Investments by E & P Companies in Renewable Energy Sources for Future Energy Needs". en. In: OnePetro. DOI: 10.2523/IPTC-19729-MS. URL: https://onepetro.org/IPTCONF/proceedings/20IPTC/1-20IPTC/D011S009R002/154555 (visited on Nov. 1, 2022).

Hatje, V., R. M. A. Pedreira, C. E. de Rezende, C. A. F. Schettini, G. C. de Souza, D. C. Marin, and P. C. Hackspacher (Sept. 2017). "The environmental impacts of one of the largest tailing dam failures worldwide". en. In: *Scientific Reports* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 10706. ISSN: 2045-2322. DOI: 10.1038/s41598-017-11143-x. URL: https://www.nature.com/articles/s41598-017-11143-x (visited on June 4, 2023).

Hunt, J. (Apr. 1992). "The Impact of the 1962 Repatriates from Algeria on the French Labor Market". en. In: *ILR Review* 45.3. Publisher: SAGE Publications Inc, pp. 556–572. ISSN: 0019-7939. DOI: 10.1177/001979399204500310. URL: https://doi.org/10.1177/001979399204500310 (visited on Mar. 19, 2024).

Jewell, S. L., G. Razzu, and C. Singleton (2020). "Who Works for Whom and the UK Gender Pay Gap". en. In: *British Journal of Industrial Relations* 58.1, pp. 50–81. ISSN: 1467-8543. DOI: 10.1111/bjir.12497. URL:

https://onlinelibrary.wiley.com/doi/abs/10.1111/bjir.12497 (visited on July 21, 2024).

Kirchberger, M. (Mar. 2017). "Natural disasters and labor markets". en. In: *Journal of Development Economics* 125, pp. 40–58. ISSN: 0304-3878. DOI: 10.1016/j.jdeveco.2016.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0304387816300943 (visited on June 3, 2023).

Kitagawa, E. M. (Dec. 1955). "Components of a Difference Between Two Rates*". In: *Journal of the American Statistical Association* 50.272, pp. 1168–1194. ISSN: 0162-1459. DOI: 10.1080/01621459.1955.10501299. URL: https://doi.org/10.1080/01621459.1955.10501299 (visited on July 30, 2024).

Kline, P., R. Saggio, and M. Sølvsten (2020). "Leave-Out Estimation of Variance Components". en. In: *Econometrica* 88.5, pp. 1859–1898. ISSN: 0012-9682. DOI: 10.3982/ECTA16410. URL: https://www.econometricsociety.org/doi/10.3982/ECTA16410 (visited on July 11, 2024).

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (Apr. 2023). "Do firm effects drift? Evidence from Washington administrative data". en. In: *Journal of Econometrics* 233.2, pp. 375–395. ISSN: 03044076. DOI: 10.1016/j.jeconom.2021.12.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304407622000604 (visited on Aug. 5, 2024).

Lavetti, K. and I. M. Schmutte (Apr. 2023). "Gender differences in sorting on wages and risk". In: *Journal of Econometrics* 233.2, pp. 507–523. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2022.06.012. URL: https://www.sciencedirect.com/science/article/pii/S030440762200183X (visited on Mar. 15, 2024).

Lebow, J. (Jan. 2022). "The labor market effects of Venezuelan migration to Colombia: reconciling conflicting results". es. In: *IZA Journal of Development and Migration* 13.1. URL: https://sciendo.com/es/article/10.2478/izajodm-2022-0005 (visited on Mar. 19, 2024).

Lechner, M. and A. Strittmatter (Feb. 2019). "Practical procedures to deal with common support problems in matching estimation". In: *Econometric Reviews* 38.2, pp. 193–207. ISSN: 0747-4938. DOI: 10.1080/07474938.2017.1318509. URL: https://doi.org/10.1080/07474938.2017.1318509 (visited on Dec. 4, 2023).

Llull, J. (July 2018). "Immigration, Wages, and Education: A Labour Market Equilibrium Structural Model". In: *The Review of Economic Studies* 85.3, pp. 1852–1896. ISSN: 0034-6527. DOI: 10.1093/restud/rdx053. URL:

https://doi.org/10.1093/restud/rdx053 (visited on Oct. 30, 2022).

Lopes, M. (Jan. 2018). *Mais de 70 mil venezuelanos entraram em Roraima em 2017*. URL: https://folhabv.com.br/noticia/CIDADES/Capital/Mais-de-70-mil-venezuelanos-entraram-em-Roraima-em-2017/35775 (visited on Oct. 29, 2022).

MacKinnon, J. G., M. Ø. Nielsen, and M. D. Webb (Feb. 2023). "Cluster-robust inference: A guide to empirical practice". en. In: *Journal of Econometrics* 232.2, pp. 272–299. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2022.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0304407622000781 (visited on June 11, 2023).

MacQueen, J. (1967). "Some Methods for classification and analysis of multivariate observations". en. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Google-Books-ID: IC4Ku_7dBFUC. URL: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992.

Manacorda, M., A. Manning, and J. Wadsworth (Feb. 2012). "The Impact of Immigration on the Structure of Wages: Theory and Evidence from Britain". In: *Journal of the European Economic Association* 10.1, pp. 120–151. ISSN: 1542-4766. DOI: 10.1111/j.1542-4774.2011.01049.x. URL: https://doi.org/10.1111/j.1542-4774.2011.01049.x (visited on Oct. 31, 2022).

Masso, J., J. Meriküll, and P. Vahter (June 2022). "The role of firms in the gender wage gap". In: *Journal of Comparative Economics* 50.2, pp. 454–473. ISSN: 0147-5967. DOI: 10.1016/j.jce.2021.10.001. URL: https://www.sciencedirect.com/science/article/pii/S0147596721000639 (visited on Mar. 1, 2024).

Maystadt, J.-F. and G. Duranton (Mar. 2019). "The development push of refugees: evidence from Tanzania". In: *Journal of Economic Geography* 19.2, pp. 299–334. ISSN: 1468-2702. DOI: 10.1093/jeg/lby020. URL: https://doi.org/10.1093/jeg/lby020 (visited on Mar. 29, 2024).

Maystadt, J.-F. and P. Verwimp (July 2014). "Winners and Losers among a Refugee-Hosting Population". In: *Economic Development and Cultural Change* 62.4. Publisher: The University of Chicago Press, pp. 769–809.

ISSN: 0013-0079. DOI: 10.1086/676458. URL: https://www.journals.uchicago.edu/doi/full/10.1086/676458 (visited on Oct. 20, 2022).

McIntosh, M. F. (May 2008). "Measuring the Labor Market Impacts of Hurricane Katrina Migration: Evidence from Houston, Texas". en. In: *American Economic Review* 98.2, pp. 54–57. ISSN: 0002-8282. DOI: 10.1257/aer.98.2.54. URL: https://www.aeaweb.org/articles?id=10.1257/aer.98.2.54 (visited on June 3, 2023).

Morales, F. and M. D. Pierola (2020). *Venezuelan migration in Peru: Short-term adjustments in the labor market*. eng. Working Paper IDB-WP-1146. IDB Working Paper Series. DOI: 10.18235/0002594. URL: https://www.econstor.eu/handle/10419/234715 (visited on Mar. 19, 2024).

Mulligan, C. B. and Y. Rubinstein (Aug. 2008). "Selection, Investment, and Women's Relative Wages Over Time*". In: *The Quarterly Journal of Economics* 123.3, pp. 1061–1110. ISSN: 0033-5533. DOI: 10.1162/qjec.2008.123.3.1061. URL: https://doi.org/10.1162/qjec.2008.123.3.1061 (visited on Sept. 30, 2024).

Oaxaca, R. (1973). "Male-Female Wage Differentials in Urban Labor Markets". In: *International Economic Review* 14.3. Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University], pp. 693–709. ISSN: 0020-6598. DOI: 10.2307/2525981. URL: https://www.jstor.org/stable/2525981 (visited on Aug. 22, 2023).

OECD (2024). *OECD Family Database - OECD*. URL: https://www.oecd.org/social/family/database.htm (visited on Mar. 1, 2024).

Oliveira, G. A. G. de (2019). *Use of the Brazilian Military Component in the Face of Venezuela's Migration Crisis*.

Ortega, F. and S. Taşpınar (July 2018). "Rising sea levels and sinking property values: Hurricane Sandy and New York's housing market". en. In: *Journal of Urban Economics* 106, pp. 81–100. ISSN: 0094-1190. DOI: 10.1016/j.jue.2018.06.005. URL: https://www.sciencedirect.com/science/article/pii/S0094119018300354 (visited on June 3, 2023).

Peri, G. and C. Sparber (July 2009). "Task Specialization, Immigration, and Wages". en. In: *American Economic Journal: Applied Economics* 1.3, pp. 135–169. ISSN: 1945-7782. DOI: 10.1257/app.1.3.135. URL: https://www.aeaweb.org/articles?id=10.1257/app.1.3.135 (visited on Oct. 31, 2022).

Ramsey, G. and G. Sánchez-Garzoli (2018). *Responding to an Exodus: Venezuela's Migration and Refugee Crisis as Seen From the Colombian and Brazilian Borders*.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". In: *Journal of the American Statistical Association* 89.427. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 846–866. ISSN: 0162-1459. DOI: 10.2307/2290910. URL: https://www.jstor.org/stable/2290910 (visited on May 23, 2023).

Ruiz, I. and C. Vargas-Silva (May 2016). "The labour market consequences of hosting refugees". en. In: *Journal of Economic Geography* 16.3, pp. 667–694. ISSN: 1468-2702, 1468-2710. DOI: 10.1093/jeg/lbv019. URL: https://academic.oup.com/joeg/article-lookup/doi/10.1093/jeg/lbv019 (visited on Oct. 20, 2022).

Ryu, H. and J. Paudel (2022). "Refugee Inflow and Labor Market Outcomes in Brazil: Evidence from the Venezuelan Exodus". en. In: *Population and Development Review* 48.1, pp. 75–96. ISSN: 1728-4457. DOI: 10.1111/padr.12452. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/padr.12452 (visited on Oct. 20, 2022).

Sant'Anna, H. and S. Shrestha (Feb. 2023). *The Effects of the Venezuelan Refugee Crisis on the Brazilian Labor Market*. arXiv:2302.04201 [econ, q-fin]. DOI: 10.48550/arXiv.2302.04201. URL: http://arxiv.org/abs/2302.04201 (visited on May 23, 2023).

Sant'Anna, P. H. C. and J. Zhao (Nov. 2020). "Doubly robust difference-in-differences estimators". en. In: *Journal of Econometrics* 219.1, pp. 101–122. ISSN: 0304-4076. DOI: 10.1016/j.jeconom.2020.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0304407620301901 (visited on May 23, 2023).

Schnoke, M., I. Lendel, J. Yochum, S. Driscoll, M. Saneda, G. Figueroa, and M. Isler (Mar. 2023). "The Economic Consequences of the East Palestine Train Derailment". In: *All Maxine Goodman Levin School of Urban Affairs Publications*, pp. 1–10. URL: https://engagedscholarship.csuohio.edu/urban_facpub/1788.

Shimer, R. and L. Smith (2000). "Assortative Matching and Search". en. In: *Econometrica* 68.2, pp. 343–369. ISSN: 1468-0262. DOI: 10.1111/1468-0262.00112. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00112 (visited on Feb. 16, 2024).

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (Feb. 2019). "Firming Up Inequality*". In: *The Quarterly Journal of Economics* 134.1,

pp. 1–50. ISSN: 0033-5533. DOI: 10.1093/qje/qjy025. URL: https://doi.org/10.1093/qje/qjy025 (visited on July 13, 2024).

Strittmatter, A. and C. Wunsch (2021). *The Gender Pay Gap Revisited with Big Data: Do Methodological Choices Matter?* en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.3798933. URL: https://papers.ssrn.com/abstract=3798933 (visited on May 23, 2023).

Taylor, J. E., M. J. Filipski, M. Alloush, A. Gupta, R. I. Rojas Valdes, and E. Gonzalez-Estrada (July 2016). "Economic impact of refugees". In: *Proceedings of the National Academy of Sciences* 113.27. Publisher: Proceedings of the National Academy of Sciences, pp. 7449–7453. DOI: 10.1073/pnas.1604566113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1604566113 (visited on Oct. 20, 2022).

Tibshirani, R., G. Walther, and T. Hastie (2001). "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63.2. Publisher: [Royal Statistical Society, Wiley], pp. 411–423. ISSN: 1369-7412. URL: https://www.jstor.org/stable/2680607 (visited on July 5, 2024).

Tumen, S. (May 2016). "The Economic Impact of Syrian Refugees on Host Countries: Quasi-experimental Evidence from Turkey". en. In: *American Economic Review* 106.5, pp. 456–460. ISSN: 0002-8282. DOI: 10.1257/aer.p20161065. URL: https://www.aeaweb.org/articles?id=10.1257/aer.p20161065 (visited on Oct. 20, 2022).

UNHCR (2022). *Venezuela situation - The UN Refugee Agency*. en. URL: https://www.unhcr.org/venezuela-emergency.html (visited on Oct. 31, 2022).

UNHCR (2023). *UNHCR Mid-year Trends 2023*. URL: https://www.unhcr.org/sites/default/files/2023-10/Mid-year-trends-2023.pdf.

Uysal, S. D. (2015). "Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling". en. In: *Journal of Applied Econometrics* 30.5, pp. 763–786. ISSN: 1099-1255. DOI: 10.1002/jae.2386. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2386 (visited on Dec. 4, 2023).

Vigdor, J. (Dec. 2008). "The Economic Aftermath of Hurricane Katrina". en. In: *Journal of Economic Perspectives* 22.4, pp. 135–154. ISSN: 0895-3309. DOI: 10.1257/jep.22.4.135. URL: https://www.aeaweb.org/articles?id=10.1257/jep.22.4.135 (visited on June 3, 2023).

Woodcock, S. D. (Aug. 2008). "Wage differentials in the presence of unobserved worker, firm, and match heterogeneity". In: *Labour Economics*. Eu-

ropean Association of Labour Economists 19th annual conference / Firms and Employees 15.4, pp. 771–793. ISSN: 0927-5371. DOI: 10.1016/j.labeco.2007.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0927537107000395 (visited on May 17, 2024).

Zissimopoulos, J. and L. A. Karoly (May 2010). "Employment and self-employment in the wake of Hurricane Katrina". In: *Demography* 47.2, pp. 345–367. ISSN: 0070-3370. DOI: 10.1353/dem.0.0099. URL: https://doi.org/10.1353/dem.0.0099 (visited on June 3, 2023).