Matching and Local Labor Market Size in Mexico

Jorge Pérez Pérez[*] Jorge Meléndez[†] José G. Nuño-Ledesma[‡]

July 18, 2025[§]

Link to most recent version

Abstract

We explore how the size of local labor markets (LLM) influences the quality of matching between workers and firms in Mexico. Using a matched employer-employee dataset comprising over 80% of all formal workers in the country, we estimate models of log wages with additive worker and workplace fixed effects, which we leverage to construct a measure of assortative matching. We find evidence of positive assortative matching between workers and firms at the aggregate level. Specifically, highly productive workers, characterized by high worker fixed effects, tend to be employed by companies with high firm fixed effects. We then correlate our matching metric to the size of LLMs. We find a positive impact of LLM size on matching in Mexico. Doubling a local labor market size increases the correlation of worker and firm fixed effects by four to seven percentage points. We also find that labor informality reduces both the strength of assortative matching and its connection to labor market size. In markets with high informality, we observe that, on average, assortative matching is negative, meaning that high-wage workers tend to work at low-wage firms. Furthermore, larger markets do not yield better worker-firm matching.

[§]For insightful comments we thank Santiago Hermo, Alan Ker, Nicolás Sidicaro, Martin Trombetta, and David Wiczer; as well as audiences at seminars hosted by Michigan State University, ITAM, LACEA, the Urban Economics Association, the Canadian Economics Society, AAEA, Banco de México, and CAES. We are grateful to Jennifer Alix-Garcia and Wolfgang Dauth for providing access to critical datasets. Karla Neri, Vicente López, Guillermo Mondragón, René Nieto, Diego Mayorga, Ariadna Martínez, Marco González, and Carolina Crispín provided great research assistance. The views and conclusions presented in this document are the exclusive responsibility of the authors and do not necessarily reflect those of Banco de México. We acknowledge financial support from Colombia Cientifica – Alianza EFI #60185 contract #FP44842-220-2018, funded by The World Bank's Scientific Ecosystems, and managed by the Colombian Ministry of Science, Technology, and Innovation. The data was accessed through the Econlab at Banco de México. The EconLab collected and processed the data as part of its effort to promote evidence-based research and foster ties between Banco de México's research staff and the academic community. Inquiries regarding the terms under which the data can be accessed should be directed to econlab@banxico.org.mx.

[*]Corresponding author. Banco de México. General Directorate of Economic Research. 18 Av. 5 de Mayo. Colonia Centro, Mexico City, México 06000. Email: jorgepp@banxico.org.mx.

[†]Banco de México. General Directorate of Economic Research. 18 Av. 5 de Mayo. Colonia Centro, Mexico City, México 06000. Email: jorge.melendez@banxico.org.mx.

[‡]University of Guelph. Department of Food, Agricultural and Resource Economics. J.D. MacLachlan Building 50 Stone Road East Guelph, Ontario, Canada N1G 2W1. Email: jnuno@uoguelph.ca.

Keywords: Assortative Matching; City size; Local Labor Markets; Mexico; Wage determination, Informality

JEL Codes: J21, J31, O54, R23

1 Introduction

Economists have long hypothesized that larger local labor markets (LLMs) facilitate better matching between workers and employers, and extensive empirical research supports the existence of these agglomeration economies (Moretti and Yi, 2024; Dauth et al., 2022; D'Costa and Overman, 2014; Duranton and Puga, 2004). Intuitively, the better matching process fuels a virtuous cycle wherein better matching enhances productivity, leading to higher salaries, which attract more skilled candidates, thereby increasing the local population (Ahlfeldt and Pietrostefani, 2019). Agglomeration economies significantly impact labor markets, driving wage gaps across LLMs and contributing more to overall earnings inequality across regions than differences within the same LLM (Dauth et al., 2022; Combes and Gobillon, 2015; D'Costa and Overman, 2014; Baum-Snow and Pavan, 2012; Gould, 2007; Duranton and Puga, 2004). Thus, a thorough understanding of any country's labor market dynamics must include a careful analysis of how enhanced matching due to larger labor markets contributes to overall wage dynamics. This study extends existing research by analyzing the relationship between local labor market size and assortative matching in Mexico, a country with structural features that differ sharply from those of more advanced economies.

To better illustrate the contribution of our work, consider that agglomeration economies may exhibit different dynamics in developing economies for several reasons. For example, while wage elasticities with respect to urban density in developing economies are comparable to those observed in developed ones, these gains are often insufficient to outweigh disutilities such as pollution, crime, and unreliable travel times (Akbar et al., 2023; Grover et al., 2023). Similarly, poor quality of transport infrastructure increases urban fragmentation and reduces the scope of agglomeration economies (Baum-Snow et al., 2017; Ghani et al., 2016). In addition, the determinants of wage dispersion in developing economies are different than those in developed economies, with workplace-level wage determinants being significantly more important in explaining wage differentials (Bassier, 2023; Diallo et al., 2022; Frías et al., 2022; Pérez Pérez and Nuño-Ledesma, 2024). Lastly, low productivity in informal markets can hinder urban development and discourage workers from improving their skills, limiting the agglomeration benefits of larger LLM (Jedwab et al., 2022; Duranton, 2015); a point which is particularly relevant to the Mexican case, where both a formal and a large informal sector co-exist.

To evaluate the relationship between LLM size and worker-workplace sorting, we first construct a metric of assortative matching. To this end, we leverage data from social security records comprising the near-universe of formal workers in Mexico from 2004 to 2018. With these records, we estimate models of log wages with additive worker and workplace fixed effects using the approach commonly referred to as AKM, first popularized by Abowd et al. (1999). To proxy assortative matching, we use the covariance of worker and workplace fixed effects at the local labor market level, following Card et al. (2013) and Dauth et al. (2022).

We proceed by taking our measure of assortative matching and correlating it with population at the local labor market level, defining local labor markets as commuting zones or commuting-zone-industry cells. To provide context for our estimations, we compare them to equivalent metrics obtained from research on labor markets in Germany. We find that the association between LLM size and the degree of positive assortative matching in formal labor markets is similar in Mexico. At the commuting-zone level and after correcting for limited mobility bias in the AKM model, we estimate the average covariance between worker and workplace fixed effects in Mexico is 0.197, compared to 0.164 in Germany (Dauth et al., 2022). Also, at this level, the slope of the relationship between log LLM size and the correlation coefficient between workplace and worker fixed effect estimates ranges from 0.04 to 0.07, close to the slope of 0.061 estimated by Dauth et al. (2022).

Although the overall extent of matching externalities looks similar in Mexico relative to what has been previously found for Germany, the features of Mexico's labor markets may lead to differences in assortative matching. We examine the relationship of several characteristics of the Mexican economy on the extent of assortative matching and its relationship to city size. Mexico is characterized by high labor informality, a low level of schooling, a preponderance of small firms and an industrial composition less geared towards services relative to Germany. We analyze all these mechanisms and find that labor informality plays an outsize role on the determination of assortative matching. In markets with high labor informality, assortative matching in formal labor markets is usually negative, meaning that high-fixed-effect workers tend to work in low-fixed-effect firms. Moreover, we find that informality hinders the matching advantages of large labor markets: among the sample of high informality labor markets, the larger ones do not have better matching between workers and firms relative to the smaller ones. We also find that agglomeration externalities from local labor market size are more prevalent in large firms and in firms in the

service sector. Markets where the population is more educated also have better matching, but this relationship dissapears once we control for informality.

We contribute to three strands of literature. First, we provide an additional example to the set of studies documenting the presence of city-size wage gaps not only in developed countries, as discussed by Baum-Snow and Pavan (2012); Gould (2007); D'Costa and Overman (2014), and De la Roca and Puga (2016), but also in developing economies (Chauvin et al., 2017; Combes et al., 2020; De la Roca et al., 2023; Duranton, 2016). Our estimates for Mexico, a middle-income country with high labor informality, validate these previous estimates. Second, we contribute to the literature on matching in labor markets and agglomeration forces (Andersson et al., 2007; Baum-Snow and Pavan, 2012; Behrens et al., 2014; Dauth et al., 2022), by showing that agglomeration forces are similar in Mexico and that labor informality reduces positive assortative matching in formal labor markets. In doing so, we also contribute to the literature on informal labor markets Ulyssea (2018) and their relationship with formal ones (Levy Algazi, 2018; Ulyssea, 2010). Lastly, we contribute to the small but growing set of studies, such as Frías et al. (2022) and Pérez Pérez and Nuño-Ledesma (2024), that adapt multi-dimensional fixed effects models to the Mexican case.

The rest of the paper proceeds as follows. Section 2 provides an overview of the data and the Mexican context. Section 3 details our AKM model estimation. Section 4 estimates the relationship between assortative matching and city and local-labor-market-level covariates and estimates the relationship between matching, city size, and informality in Mexico. Section 5 discusses potential mechanisms that may change the relationship between matching and city size in Mexico. Section 6 concludes.

2 Data

We use social security records from the Mexican Social Security Institute (IMSS, by its Spanish acronym). IMSS is the Mexican government's paramount social security, pension, and public health administrator. Salaried workers in the private sector are legally required to register with IMSS. According to the government's estimate, 83% of all formal workers are affiliated with IMSS (Pérez Pérez and Nuño-Ledesma, 2024). The records at our disposal report monthly observations from November 2004 to December 2018. The data for the last month contained information for

approximately 20.1 million workers.[1] Our data are not without limitations. Public sector employees are not included in our data because their records are managed by a different agency. Wages for workers with high earnings are censored. [2] Information on persons working in the shadow economy is absent from IMSS data.[3] When estimating the AKM models, the dependent variable of interest is the natural logarithm of real daily taxable income.[4] Taxable income may include remunerations made to the worker other than wage.[5] The dataset also includes information on gender, age, and registration date to IMSS. The dataset does not include information on education or hours worked. Our primary sample of interest consists of prime-age men (25-54 years old) who have likely completed their education. Thus, their estimated worker effect should include wage variance attributable to education. Regarding employer information, the data includes workplace and economic sector identifiers.

We complement IMSS data with city-level characteristics from Mexican censuses and intercensal surveys for 2005, 2010, and 2020. We construct commuting zones level labor informality rates using data from Mexico's labor survey, ENOE. [6] Regarding local-labor-market-level characteristics, we rely on the definition of local labor markets (and the associated dataset) used by Aldeco et al. (2024), which divides Mexico into 777 local labor markets.[7]

[1] We rely on the *registro patronal* as our employer identifier because it is more precise than alternatives like the *Registro Federal de Contribuyentes*. Assigned for social security administration, the former more accurately captures employment structures in settings where multiple employers operate within a single plant. *Registros* are anonymized by our data provider, and their masking procedures are inconsequential to our estimates. See online Appendix A for details on variable construction and descriptive statistics.

[2] The censoring limit was 25 daily minimum wages before 2017 and 25 "units of measurement and update" after 2017. These thresholds amount to about 1,539 pesos per day, or approximately 80 USD at 2018 exchange rates. The presence of top-coding may bias our results. Wage censoring will reduce the correlation between worker and firm fixed effects for high-fixed-effect firms. If censoring is more common in large cities or among high-skill workers, we may underestimate matching in large cities. Only about 2% of workers have censored wages in our sample, so we might expect this bias to be small. Nevertheless, for robustness, we imputed wages for the censored observations following Card et al. (2013) and calculated the correlations between worker and firm fixed effects from AKM estimates using data that included imputed wages. The resulting firm-worker fixed-effect correlations are nearly unchanged, suggesting that our main results are robust to wage censoring. Appendix B provides details about this procedure.

[3] Informality is widespread in Mexico; according to the country's National Survey of Occupation and Employment (ENOE by its acronym in Spanish), 55% of all workers operate in the informal economy Banco de México (2023).

[4] As Dauth et al. (2022) note, using nominal or real wages only changes the scale of the firms' fixed effects.

[5] One may be concerned about firms manipulating compensation schemes and underreporting wages. Kumler et al. (2020) show that wage underreporting has declined since 1997, due to the Mexican pension reform that tied pension benefits to reported wages. Additionally, Puggioni et al. (2022) argue that underreported wages should not be a problem for the period they studied.

[6] For the analysis at the commuting zone and commuting-zone-sector levels, we estimate informality using census data, following Aldeco et al. (2024). This ensures that the measurement of informality is appropriate at each level of aggregation we consider.

[7] Aldeco et al. (2024) calculate commuting zones using the methodology in Fowler and Jensen (2020). Appendix

3 Constructing a Measure of Assortative Matching

Here we describe the methodology we used to construct our measure of assortative matching. Specifically, we measure assortative matching as the correlation between worker and workplace fixed effects, estimated using log-wage models following the AKM framework Abowd et al. (1999), as in Card et al. (2013, 2018). In AKM models, log wages are modeled as a function of additive worker and workplace fixed effects:

$$\ln\left(wage_{it}\right) = \alpha_i + \psi_{J(i,t)} + X'_{it}\beta + r_{it}. \tag{1}$$

Here, $wage_{it}$ is the real wage of worker i at time t. The vector of fixed effects $\alpha_{\mathbf{i}}$ captures the influence of all time-invariant worker characteristics. Similarly, the vector of fixed effects $\psi_{J(i,t)}$ collects time-invariant factors at the workplace level for workplace J where worker i was employed at time t. The vector X_{it} includes control variables, which in our estimations include functions of age and time-interval trends. We estimate equation (1) by OLS with a preconditioned iterative gradient method (Card et al., 2013).[8] We generate estimates of the model for three discrete time segments: 2004-2008, 2009-2013, and 2014-2018.[9]Generating results for three time segments allows us to better describe potential changes in the total-variance contributions made over time by each of the components because the set of attributes that remain time invariant in any given panel can change with the length of the panel, as pointed out by Millimet and Bellemare (2025). In previous work, Pérez Pérez and Nuño-Ledesma (2024) show that the variance contribution to wage inequality of workplace effects has increased over time, while the contribution of worker effects has decreased between 2004 and 2018.

In AKM models, worker mobility identifies firm and worker effects. As pointed out by Abowd et al. (1999), these effects can be disentangled by worker mobility –generated when workers

C provides details about the commuting zone construction.

[8]We provide validation exercises for the linearity in equation (1) and the uncorrelatedness of the error term r_{it} and the fixed effects and covariates in online Appendix D. Specifically, we follow Card et al. (2013) and show that firms are exchangeable: for a worker, moving from firm A to firm B has approximately the same effect on wages as moving from firm B to firm A. Pérez Pérez and Nuño-Ledesma (2024) also provide evidence of the validity of AKM models in this sample.

[9]Our periods cover fewer years than those in Dauth et al. (2022). However, identification issues are not a primary concern since we use monthly frequency data. Additionally, we conduct our regression analysis of city size and matching by pooling the periods and including period fixed effects.

change employers– that creates a network of directly or indirectly connected workplaces. We restrict estimation to the largest "connected set" of workplaces in each time interval across which workers change jobs at least once (Abowd et al., 1999).[10]

Table 1 provides summary statistics for the AKM model outcomes. The models account for approximately 94% of the variance in log wages across all periods. The table also shows an increasing role of assortative matching in explaining wage variations in Mexico, with the correlation rising from 0.21 during 2004–2008 to 0.26 in 2014–2018. Nevertheless, this correlation is still far below the 0.64 observed in Germany during the 2008–2014 period.

We follow Card et al. (2013) and use our estimated model to decompose the variance of wages into the shares attributed to each component:

$$\begin{aligned}\text{Var}\left(\ln \text{wage}_{it}\right) = &\underbrace{\text{Var}(\alpha_i)}_{\text{workers}} + \underbrace{\text{Var}(\psi_{\mathbf{J}(i,t)})}_{\text{workplaces}} + \text{Var}(x_{it}'\beta) + \text{Var}(r_{it}) \\ &+ 2\underbrace{\text{Cov}(\alpha_i, \psi_{\mathbf{J}(i,t)})}_{\text{Assortative Matching}} + 2\text{Cov}(\psi_{\mathbf{J}(i,t)}, x_{it}'\beta) + 2\text{Cov}(\alpha_i, x_{it}'\beta).\end{aligned} \tag{2}$$

The last rows of Table 1 show the results of this decomposition. Worker effects explained about 44% of wage variance in 2004-2008, and their contribution to total variance has decreased over time. In contrast, the contribution of workplace effects and sorting has been increasing over time. Sorting explains 16% of wage variance in 2004-2008, whereas it explains approximately 19% of wage variance in 2014-2018. This pattern suggests that assortative matching is increasingly important in determining wage variance in Mexico over time.[11]

[10] Online appendix table A.1 shows that descriptive statistics are similar in the full and connected set samples.

[11] A pervasive issue in these decompositions is the bias in the plug-in OLS estimates of the variance components of equation (2). Kline et al. (2020) show that even if OLS estimates of the fixed effects are unbiased, estimates of their variance may be biased, usually attenuating estimates of the covariance between worker and workplace fixed effects. Bonhomme et al. (2019) show that the bias in these variance components worsens when there is limited mobility of workers across workplaces. To address these issues, in Appendix section E, we repeat the decomposition using alternative estimators for the variance components of equation 2. The results of these decomposition exercises with limited mobility bias corrections are similar to the baseline results.

Table 1: AKM Model Estimation Results

	Interval 1 2004-2008	Interval 2 2009-2013	Interval 3 2014-2018
Worker and workplace parameters			
Number of worker effects	11,363,073	13,083,589	15,512,438
Number of workplace effects	858,480	892,929	1,009,320
Summary of parameter estimates			
St. dev. of worker effects	0.539	0.520	0.503
St. dev. of workplace effects	0.463	0.493	0.503
Correlation worker/workplace effects	0.208	0.226	0.262
Correlation worker effects/Xb	-0.079	-0.034	-0.067
Correlation workplace effects/Xb	-0.002	0.008	0.003
Goodness of fit			
St. dev. of log wages	0.808	0.823	0.829
RMSE	0.195	0.198	0.200
R Squared	0.942	0.942	0.942
Adj. R Squared	0.939	0.940	0.940
Total variance shares			
Worker effects	0.444	0.398	0.369
Workplace effects	0.328	0.359	0.369
2Cov(worker effects, workplace effects)	0.159	0.171	0.193
Remainder	0.069	0.072	0.069
Total variance shares corrected for limited mobility bias			
Worker effects	0.434	0.390	0.366
Workplace effects	0.247	0.280	0.291
2Cov(worker effects, workplace effects)	0.238	0.246	0.261
Remainder	0.081	0.084	0.082

Source: Authors' calculations using IMSS data. Results from estimating equation (1) via OLS with a pre-conditioned gradient method following Card et al. (2013). Total variance shares results from equation (2). Estimations are restricted to prime-aged men (ages 25-54) in the largest connected set per time interval. All the estimations include the following controls: age, age squared, age cube, and a monthly time trend. RMSE is the root mean squared error. Results corrected for limited mobility bias are obtained using the methodology of Bonhomme et al. (2019) using five firm clusters.

4 Assortative Matching and Labor Market Size

We proceed to analyze how the strength of positive assortative matching varies with local labor market size across Mexico's LLMs. To this end, we regress our measure of assortative matching against the logarithm of the population at the LLM level, following Dauth et al. (2022).[12]. We conduct our analysis at the commuting zone level using commuting zones defined by Aldeco et al. (2024). To delimit the analysis to specific labor markets further, we also conduct our analysis at the commuting zone-industry level.[13]To contextualize our findings, we compare our estimates to previous estimations of equivalent metrics obtained for Germany by Dauth et al. (2022). Each panel in Figure 1, shows a scatterplot depicting the association between log-population and our matching metric at two levels of aggregation: commuting zone and commuting-zone-industry. The relationship aligns with the German estimates.

Further econometric analysis confirms that the link between market size and assortative matching is similar in Mexico relative to Germany at the commuting zone level. We pool the estimated correlation coefficients between worker and workplace fixed effects in a single AKM model and include binary variables indicating the period accounted for in the regression.[14] We show the results of our estimations in Table 2. Panel A, column 1 shows that the correlation between labor market size and assortative matching for Mexico's commuting zones is positive and statistically significant. The link between matching and labor market size is similar to that found in Dauth et al. (2022) for Germany. Doubling the commuting-zone population increases the correlation coefficient between worker and workplace fixed effects by 6.7 p.p. Column 2 shows estimates relying on commuting-zone-industry cells as the definition of LLM.[15] At this level, the estimated association between population and matching is a bit weaker. Doubling the population size of a commuting zone-industry cell increases assortative matching by 5.3 p.p.

In the remaining panels, B to D of Table 2, we show estimates from alternative specifications,

[12]In Appendix Table F.1, we show that we obtain similar results using log-employment as a proxy variable of labor market size.

[13]We use a two-digit NAICS industry classification. Our data includes 22 industries. For the estimates at the commuting zone by industry level, we exclude cells with fewer than 50 workers or fewer than five firms.

[14]Appendix table F.2 shows that we obtain similar results if we estimate these regressions for each time interval separately.

[15]For the commuting zone by industry estimates, we restrict to cells with more than five firms and more than 50 workers, following Dauth et al. (2022).

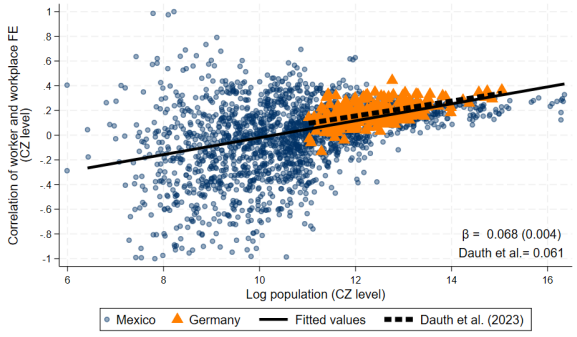
Table 2: City Size and Assortative Matching in Mexico's Formal Labor Markets

Dependent variable: correlation of worker and workplace FE		
	(1)	(2)
	CZ	CZ-Industry
A: Baseline Model		
Log Population	0.0679***	0.0531***
	(0.004)	(0.004)
R^2	0.155	0.090
B: Correlation of worker and residual workplace FE		
Log Population	0.0640***	0.0529***
	(0.004)	(0.004)
R^2	0.146	0.090
C: Log population instrumented with population in 1921-1950		
Log Population	0.0592**	0.0420***
	(0.009)	(0.013)
R^2	0.152	0.086
First-stage F	174.786	4.757
N for Panels A-C	1,961	10,118
Mean of dep. variable	-0.014	0.000
D: Corrected for limited mobility bias		
Log Population	0.0408***	0.0406***
	(0.004)	(0.003)
R^2	0.084	0.070
N	1,926	10,101
Mean of dep. variable	0.197	0.182
E: Dropping the 10% largest and smallest areas		
Log Population	0.0399***	0.0605***
	(0.006)	(0.005)
R^2	0.139	0.075
N	588	7,897
Mean of dep. variable	0.131	0.024

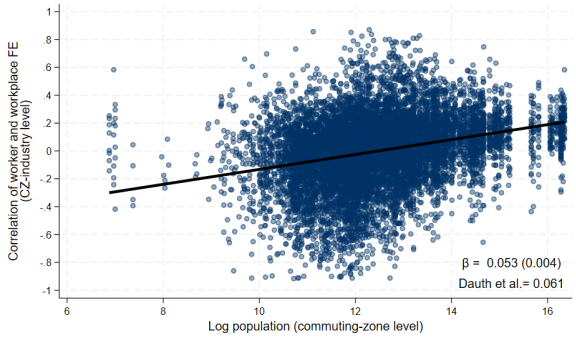
Source: Author's calculations using IMSS and INEGI data. "CZ" stands for commuting zone. The regressions pool data from 2004-2008, 2009-2013, and 2014-2018 with interval dummies. Column (2) restricts to cells with over five firms and 50 workers. Panels show the following estimates: A - baseline; B - industry-demeaned workplace fixed effects; C - we instrument population with historical population at the CZ level, relying on historical population estimates from Alix-Garcia and Sellars (2020). Log population instrumented with log population in 1921, 1930, 1940, and 1950 (Table shows different specifications of the historical population IV); D - Bonhomme et al. (2019)'s limited-mobility bias correction with five workplace clusters, and E - excluding extreme populations. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

Figure 1: City Size and Assortative Matching in Mexico's Formal Labor Markets

(a) Commuting Zone



(b) Commuting Zone - Industry



Source: Author's calculations using IMSS and INEGI data. Each panel displays a scatterplot illustrating the relationship between log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone and commuting zone industry levels. For comparison, panel (a) displays the relationship estimated for Germany by Dauth et al. (2022). For panel (b), we restrict to cells with more than five firms and more than 50 workers. We classify industries according to a 2-digit NAICS classification. The bottom-right values display the slope of a linear regression corresponding to the displayed relationship. The regressions include dummies for each time interval. Clustered standard errors at the commuting-zone level in parentheses.

which we conducted to gauge the robustness of our findings. To residualize firm fixed effects for Panel B, we regress them on industrial sector effects (NAICS 2-digit level). We then calculate the correlation between the residuals of the previous regression and the person fixed effects at each geographical level. By doing this, the estimated correlations between city size and matching control for different industry compositions across cities. Using these "residual" fixed effects, the association between city size and matching is weaker than in previous panels. This pattern of results indicates that part of the effect seen in the estimates with unadjusted fixed effects was due to industries with better labor market matching located in larger cities.

To avoid reverse causality between matching and demographic growth, we instrument current population with population in 1921, 1930, 1940, and 1950 as reported by Alix-Garcia and Sellars (2020). These historical estimates are based on 15 by 15-kilometer grid cells, which we aggregate to the desired levels. Panel C shows that this method results in a weaker matching-market size link compared to Panel A. [16]

[16] We provide details about historical population and data in Appendix H, and we show estimates instrumenting with historical population values from different years between 1921 and 1950 in Appendix Table H.1. More recent values

In Panel D, we report results from an econometric model adjusting for limited mobility bias as proposed by Bonhomme et al. (2019).[17] For these estimates, we cluster firms into five clusters using 20 percentiles of the within-firm distribution of wages as clustering variables. Then, we re-estimate the AKM model with this reduced number of workplace-effect parameters and recalculate the correlations between worker and workplace effects. The correction reduces the magnitude of the correlation, which may suggest that limited mobility had inflated our estimates. However, because the correction is relatively modest, and the corrected estimates still show a statistically significant positive association between LLM size and matching, we interpret this as evidence that limited mobility does not fundamentally alter the pattern we find.

Last, Panel E removes the largest and smallest areas from the sample. The results are similar in this restricted sample, implying the small association we find is not due to the smallest commuting zones or to Mexico City's influence.[18] Overall, the city-size advantage for matching in the labor market in Mexico is similar to the results of Dauth et al. (2022). We now explore frictions that weaken LLM-size matching externalities in Mexico.

5 Factors Limiting Matching Externalities from Larger Labor Markets in Mexico

We now assess potential frictions that may mediate agglomeration externalities in Mexico. At the local labor market level, we focus on the role of labor informality and differences in educa-

of historical populations are stronger predictors of current population. We have historical population data for 1900 and 1910. However, we excluded these years in our estimations because the results of the Hansen test for overidentifying restrictions, when including these periods, rejected the validity of the instruments.

[17] We use the Bonhomme et al. (2019) correction for two reasons. First, we diagnosed the bias in our estimation of the variance of workplace fixed effects using the method from Jochmans and Weidner (2019). We calculated the sum of the trace of the normalized Laplacian matrix for the connected set. The estimated bias in percentage terms is around 25.3%, 23.5%, and 23.8% for 2004-2008, 2009-2013, and 2014-2018, respectively. When using the Bonhomme et al. (2019) methodology, our estimates are 24.3%, 21.8%, and 20.9% smaller than the uncorrected estimates of workplace effects variance, close to the estimated bias obtained with the Jochmans and Weidner (2019) method. Second, the Bonhomme et al. (2019) correction is computationally more feasible than the correction proposed by Kline et al. (2020).

[18] Since the correlation coefficients between worker and firm fixed effects are themseleves estimated, there may be heteroskedasticity in the regression model because in smaller local labor markets the estimated correlation coefficient will have a larger sampling error. To check if this heteroskedasticity affects our results, we re-estimate the regressions in Table 2 using a non-parametric correction for heteroskedasticity (Cadena, 2014). The results are in Appendix Table F.3, and are similar to our baseline results.

tional attainment. At the firm level, we focus on firm size and on differences in matching across industries.[19]

We choose these mechanisms since we deem them as the most distinctive of Mexican labor markets relative to those in developed countries. Nevertheless, other mechanisms may also influence the extent of matching and its relationship to city size. For example, mobility frictions and spatial mismatch may prevent workers from reaching firms that are a good match. Information asymmetries may also prevent good matches when firms are unable to judge the quality of their workers.

5.1 Labor Informality

Labor market informality is high in Mexico. According to Mexico's labor survey (ENOE), 54% of employed workers were informal as of April 2025. Informal workers may not pay income taxes, lack labor stability, and do not have access to social security through their employers. The high rate of labor informality has been documented as a limitation to productivity growth in Mexico (Levy Algazi, 2018).

We hypothesize that a large informal labor market may depress assortative matching in the formal labor market for multiple reasons, which come from the study of agglomeration effects. The first reason may be increased search costs for firms looking for workers who constitute a good match. The reduction in search costs as economies grow larger is one of the traditional channels for agglomeration effects associated with city size (Henderson, 1986). However, informality may disrupt this relationship by attracting workers to informal jobs, decreasing matches, and increasing search costs for formal labor market firms (Helsley and Strange, 1990). A larger share of informal firms may make it harder to fill vacancies in the formal sector if it captures a large share of the labor supply (Petrongolo and Pissarides, 2006), or if it increases reservation wages for workers who use informal wages to set reference wages against which to compare formal job offers.

Informality may depress assortative matching and city-size advantages by weakening the benefits of acquiring specialized human capital that leads to matches in formal firms. Rotemberg and

[19] We also analyze the role of unions in Appendix section J. Unionization rates are only available at the metropolitan area level; therefore, our analysis of the role of unions is conducted at the metro-area level, rather than the commuting-zone level. Because of this lack of comparability, we choose not to emphasize it here.

Saloner (2000) argue that for workers to engage in specific human capital acquisition, they need to be able to recoup the cost of the investment afterwards. While a larger availability of informal jobs may allow workers to experiment before specializing (Wheeler, 2008; Bleakley and Lin, 2012), it may also deter them from specializing. Acquiring specific human capital may not be profitable for workers in markets where informal labor opportunities that do not require specific human capital are abundant.

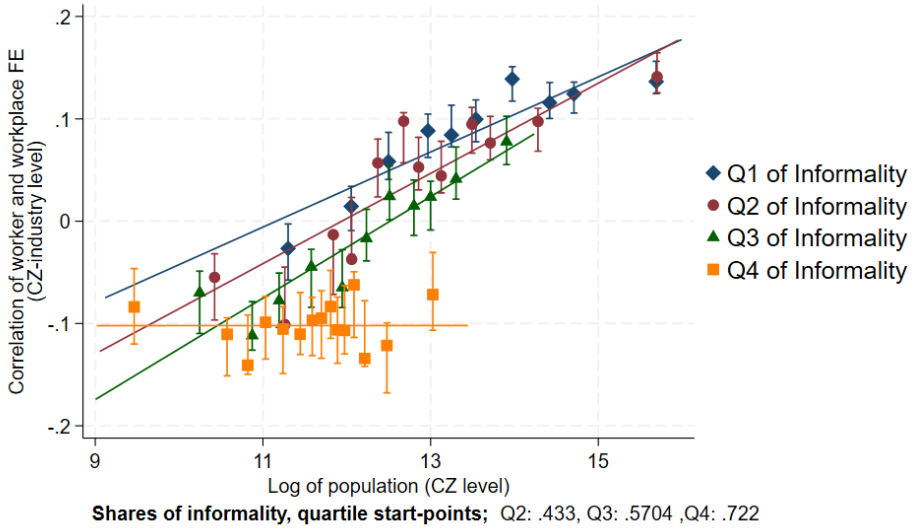
High labor informality may also weaken the incentives for formal firms to find good worker matches. If good matches complement production and the presence of informal firms makes formal firms less productive, then, even in large markets, formal firms will form fewer good matches in markets with high informality. Formal firms may also decide to hire informal workers instead of formal workers, further weakening matching in the formal sector.

Figure 2 shows the relationship between city size and assortative matching, separating the sample by quantiles of informality, using informality data at the municipality level from the economic censuses as compiled by Aldeco et al. (2024). The Figure shows that informality weakens assortative matching in the formal sector and erases the advantages of large markets for matching. For markets in the first three quartiles of informality, with informality rates below 72%, the relationship is similar to that in Figure 1. For the last quartile, which has the highest levels of informality, the relationship is substantially different. On average, in these markets, assortative matching is negative: the average correlation between worker and workplace fixed effects, weighted by employment, is -0.001, and this negative correlation holds across all values of the populations of the commuting zones. Furthermore, assortative matching does not appear to increase linearly with city size, as observed in the previous three quartiles; instead, the relationship remains flat, except for the largest markets.[20]

We confirm this relationship by estimating the slope of the relationship between labor market size and matching separately by quartiles of labor informality in column (1) of Table 3. The slope for the first three quartiles of informality is positive and significant, while it is close to zero and precisely estimated for the markets with higher informality.

[20]The lower correlation of worker and firm fixed effects for markets with higher informality and the flat relationship of assortative matching and city size in these markets also appears at the commuting zone level (Appendix Figure I.1) and across periods (Appendix Figure I.2). We also observe a lower correlation among high informality commuting zones if we exclude the southern region of Mexico from the estimation (Appendix Figure I.3).

Figure 2: City Size and Assortative Matching in Mexico's Formal Labor Markets. Commuting Zone-Industry by Informality Quartiles



Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry level for each quartile of the informality rate. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

Table 3: City Size and Assortative Matching in Mexico's Formal Labor Markets: Controlling for Informality and Years of Schooling

Dependent variable: correlation of worker and workplace FE				
			Joint estimation	
	(1)	(2)	(3)	(4)
	Informality rate	Mean years of schooling	Informality rate	Mean years of schooling
	CZ-Industry	CZ-Industry	CZ-Industry	CZ-Industry
Log Pop*first quartile	0.0366***	0.0255**	0.0404***	
	(0.006)	(0.013)	(0.010)	
Log Pop*second quartile	0.0440***	0.0508***	0.0452***	0.0089
	(0.005)	(0.005)	(0.009)	(0.100)
Log Pop*third quartile	0.0496***	0.0453***	0.0459***	-0.0010
	(0.007)	(0.006)	(0.009)	(0.115)
Log Pop*fourth quartile	0.0004	0.0348***	-0.0024	-0.0082
	(0.012)	(0.004)	(0.013)	(0.113)
Dummy first quartile	-0.4326**	-0.3857**	-0.4948***	
	(0.084)	(0.152)	(0.129)	
Dummy second quartile	-0.5476***	-0.6338***	-0.5738***	-0.0922
	(0.062)	(0.064)	(0.108)	(0.100)
Dummy third quartile	-0.6456***	-0.5522***	-0.6080***	0.0276
	(0.089)	(0.078)	(0.113)	(0.115)
Dummy fourth quartile	-0.1306	-0.3789***	-0.1010	0.1264
	(0.140)	(0.058)	(0.152)	(0.113)
Mean of dep. variable	0.0002	0.0002	0.0002	0.0002
R^2	0.117	0.103	0.118	0.118
Obs.	10,117	10,117	10,117	10,117

Source: Author's calculations using IMSS and INEGI data. The columns display the results of regressions of the correlation coefficient between worker and workplace effects from AKM model estimates and log population, interacted with informality rates and mean years of schooling quartiles at the commuting-zone-industry level. CZ stands for commuting zone. We restrict to cells with more than five firms and more than 50 workers. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

5.2 Schooling

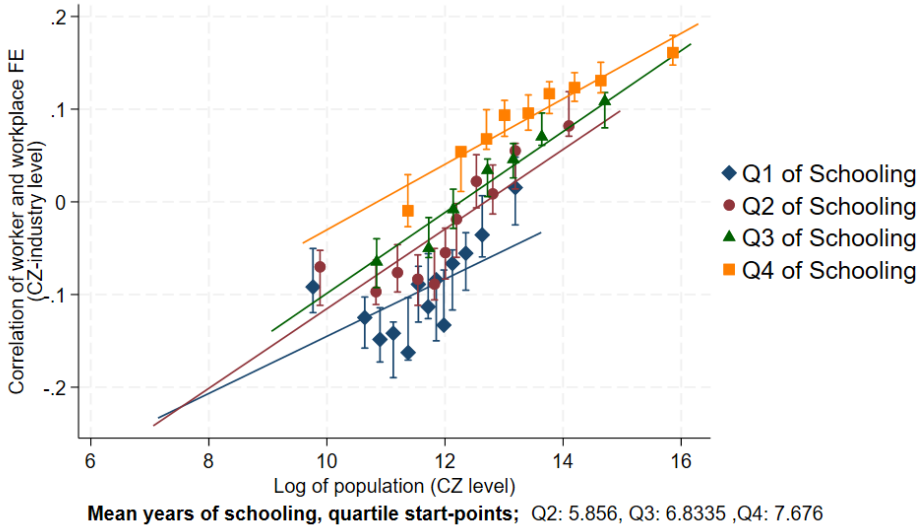
Schooling levels may modify both the level and the slope of the relationship between city size and matching. Individuals with higher levels of education may be better informed about vacancies, the skill requirements of each job, and the productivity of particular firms. All of these reasons may promote better matching for highly educated workers. As for the relationship with city size, part of the advantage of large cities in generating better matches between employers and employees may occur because more educated workers sort to bigger cities, and at the same time, firms in big cities demand workers with higher education. If so, we would expect the relationship between city size and matching to be steeper among places where individuals are more educated.

The evidence on the role of schooling in the city-size wage premium has been mixed. De la Roca and Puga (2016) do not find differences in the return to large-city experience between high- and low-education workers. However, they do observe differences in the returns to city size for high- and low-ability workers, measuring ability using worker fixed effects from a wage model. Bacolod et al. (2023) finds that high-learning-ability individuals sort to big cities. This sorting may influence both matching and education in larger cities. Eckert et al. (2022) show that the returns to big cities for migrants are larger for those with higher education, who find it easier to integrate into large-cities labor markets.

Figure 3 shows that positive assortative matching is strongest in commuting zones with higher average schooling and that the advantage of large cities in matching is also higher as average education is higher in each labor market. For commuting zones in the top three schooling quartiles (with an average of more than 5.8 years of schooling), the correlation between worker and workplace fixed effects is higher, and it increases steeply with population, reaching nearly 0.2 in the largest commuting zones. By contrast, in low schooling markets (with an average of fewer than 5.8 years of schooling), the relationship is flatter.

This pattern suggests that schooling is an enabling condition for positive assortative matching. More educated local labor markets may increase returns to skill complementarity and improve the efficiency of job search. These effects reinforce agglomeration benefits, amplifying matching quality. The evidence supports the view that schooling not only drives productivity directly but also conditions the market's ability to generate productive matches.

Figure 3: City Size and Assortative Matching in Mexico's Formal Labor Markets. Commuting Zone-Industry by Years of Schooling Quartiles



Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry level for each quartile of the mean years of schooling. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

The patterns in Figure 3 are confirmed in column (2) of Table 3, which shows a robust positive interaction between city size and average years of schooling on assortative matching. The estimated coefficients on log population are increasing across schooling quartiles and are statistically significant in all. Notably, the interaction effect for the bottom quartile is smaller and only marginally significant, consistent with a flatter, weaker relationship at low levels of schooling. This supports the interpretation that education enables agglomeration economies through worker-firm matching.[21]

5.3 Firm Size and Firm Industry Composition

We now examine the role of features of Mexican firms that distinguish them from those in other countries: their industry and size. Compared to Germany, where 71% of employment is in services (World Bank, 2023), Mexico has a larger share of manufacturing employment and employment in other sectors. At the same time, Mexico has a large share of small firms: Levy Algazi (2018) estimates that out of 6.73 million firms in Mexico in 2013, 6.3 million had five workers or fewer.

We first examine whether agglomeration externalities are stronger in markets with a larger share of workers in the service sector. Such a difference could naturally arise due to the higher job turnover that service industries tend to experience.[22] Figure 4a shows the relationship between city size and matching among service and non-service firms. This relationship is smaller for firms outside the service industries, but the difference in slopes is not statistically significant.

We further explore if Mexico's industry composition may lead to a different degree of matching externalities of city size through a counterfactual exercise, where we estimate the relationship between city size and matching in Mexico if it had the same industrial composition as Germany. This exercise helps illustrate if agglomeration externalities are indeed weaker in Mexico. For this analysis, we recalculate the correlation between worker and firm FE at the commuting-zone level, reweighting the individual observations to weight Mexican industries according to their employment share for Germany.[23]

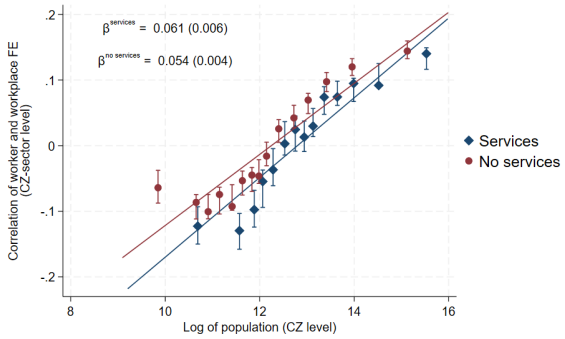
[21] We examine this relationship at the commuting zone level in Appendix Figure I.4. At this level, positive assortative matching remains weaker in low-education areas. We also show estimates by time interval in Appendix Figure I.5. The analysis by period shows the same patterns, although the estimated binned scatterplots are noisier.

[22] Box 4 in Banco de México (2020) documents higher turnover in service industries relative to other industries in Mexico.

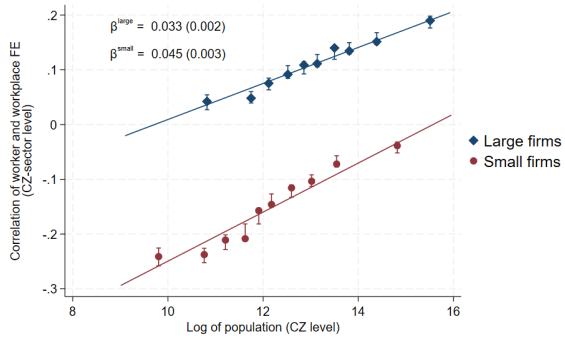
[23] We obtain data for Germany's industrial composition for 2008-2018 from Eurostat. Then, we match industries

Figure 4: City Size and Assortative Matching in Mexico's Formal Labor Markets: Industry and Firm Size Differences

(a) Industry



(b) Firm size



Source: Author's calculations using IMSS and INEGI data. Each panel displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone level. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. Panel (a) illustrates the relationship for service and non-service industries, respectively. We classify industries according to a 2-digit NAICS classification. Panel (b) shows the relationship separating large firms (16 or more workers) and small firms (fewer than 16 workers). The top left values display the slope of a linear regression corresponding to the displayed relationship. The regression includes dummies for each time interval. Standard errors in parentheses. To produce the scatter plots, we used the `binsreg` command (Cattaneo et al., 2024a,b) with the default settings. Clustered standard errors at the commuting-zone level in parentheses.

Figure 5 shows the results of this exercise. Panel (a) illustrates the relationship between population and assortative matching at the commuting zone level in Mexico, restricting the sample to observations where data on Germany's industry composition is available. The estimated relationship is slightly steeper than that of Table 2, Panel A, column (1), using the whole sample. Panel (b) shows the relationship after reweighting the sample to match Germany's industrial composition. The estimated relationship is 0.077, approximately one-third larger than that estimated for Germany by Dauth et al. (2022). This larger slope suggests that Mexico's economy has a larger share of employment in industries where matching externalities due to labor market size are weaker. If Mexico were to have the same industrial composition as Germany, these matching externalities would be stronger.[24]

Figure 4b estimates the strength of assortative matching and its relationship to labor market size, categorizing firms into two groups: large firms (with 16 workers or more) and small firms (with fewer than 16 workers). The differences are striking: small firms exhibit negative assortative matching on average across all values of the commuting zone size distribution, whereas large firms exhibit positive assortative matching. This is consistent with a search costs channel: smaller firms may be unable to bear the search costs associated with finding better matches (Henderson, 1986; Helsley and Strange, 1990). Small firms appear to benefit more from larger city sizes than large firms in terms of enhancing their ability to match with better workers. This pattern would be consistent with smaller firms experiencing bigger benefits from reduced search costs in larger labor markets.

at the NACE level to their counterparts at the NAICS level to obtain German employment shares at the NAICS level. We exclude the first interval (2004-2008) of data because industrial composition data for Germany were unavailable. Furthermore, we are unable to match the following NAICS sectors to NACE data: educational services, health care and social assistance, other services (except public administration), and public administration.

To obtain correlations between worker and firm effects at the commuting-zone level, reweighting to match Germany's industrial composition, we first standardize the worker and firm FE. Then, we estimate:

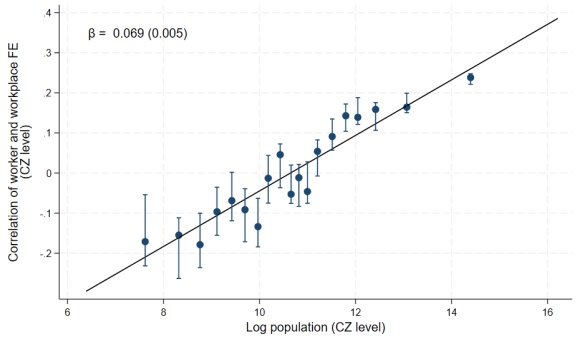
$$\text{worker FE}_{ijt} = \alpha_{c(jt)} CZ_{c(jt)} + \beta_{c(jt)} CZ_{c(jt)} \text{firm FE}_{jt} + \delta_t + \varepsilon_{ijt},$$

weighting each observation by $w_{ij} = \theta^G_{s(j)t} / N^{MX}_{s(j)}$, the share of employment for sector $s(j)$ in Germany divided by the employment in that sector in Mexico. In the equation, $CZ_{c(jt)}$ are indicators for the commuting zone of firm j at time t. The coefficients $\beta_{c(jt)}$ are the correlations between worker FE and firm FE at the commuting-zone level. We use these $\beta_{c(j)}$ coefficients as our reweighted measure of assortative matching.

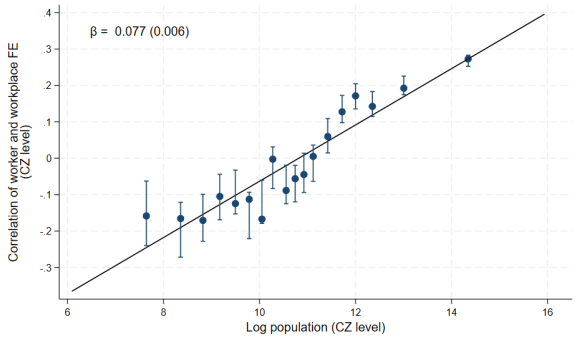
[24] We reproduce these results at the commuting zone level in Appendix Figure G.1 and Appendix Table G.1. The results at the commuting zone level are similar to those in the text. An additional concern is whether differences between Mexico and Germany arise from differences in the size of the commuting zones. In Appendix Figure G.2 and Table G.2, we show that the results are similar if we focus on commuting zones larger than 50,000 inhabitants, which are more similar to the German commuting zones in terms of size.

Figure 5: City Size and Assortative Matching in Mexico's Formal Labor Markets: Germany's industrial composition counterfactual.

(a) Mexico 2009-2018 restricted sample



(b) Mexico 2009-2018 counterfactual



Source: Author's calculations using IMSS, INEGI, and Eurostat data. Each panel displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone level. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. Panel (a) illustrates the relationship restricting the sample to industries for which we have data on Germany's industrial composition. For Panel (b), we reweighted the sample to match Germany's industrial composition and re-estimated the correlation between worker and workplace fixed effects. To produce the scatter plots, we used the `binsreg` command (Cattaneo et al., 2024a,b) with the default settings. Clustered standard errors at the commuting-zone level in parentheses.

6 Concluding Remarks

Using administrative Mexican social security records, we provide a descriptive analysis of the country's underlying mechanisms of wage variation and worker-firm sorting patterns.

Our findings highlight a structural feature that erodes the agglomeration gains from labor market size in Mexico. Widespread informality disrupts positive assortative matching in the formal sector, possibly by diverting workers into informal employment, raising search costs and reservation wages, and discouraging human capital investment and specialization. These frictions are particularly acute in larger cities, where agglomeration effects would otherwise strengthen worker-workplace matching.

In addition, our findings suggest that agglomeration externalities, which often amplify positive assortative matching in larger cities, are nevertheless present in Mexico's labor market. Labor market policy designers will have to consider whether it is preferable to strengthen agglomeration effects or improve matching within cities. On the one hand, higher agglomeration effects may enhance productivity, especially in large urban areas. However, this process may increase inequality across cities. On the other hand, improving matching quality within cities may reduce agglomeration effects and the gap between matching quality across cities. Additionally, informality needs to be considered for policies oriented to improve matching within cities or increase agglomeration effects. Depending on the level of informality, policies aiming to improve worker-firm matching in the formal sector may have heterogeneous results.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work, the authors used ChatGPT 3.5 and ChatGPT 4.0 to check spelling, improve clarity, and format BibTeX entries. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, 67(2):251–333.

Ahlfeldt, G. M. and Pietrostefani, E. (2019). The economic effects of density: A synthesis. *Journal of Urban Economics*, 111:93–107.

Akbar, P., Couture, V., Duranton, G., and Storeygard, A. (2023). Mobility and congestion in urban india. *American Economic Review*, 113(4):1083–1111.

Aldeco, L., Calderón, M., Chiquiar, D., Hanson, G., Pérez Pérez, J., and Velázquez, C. (2024). Local labor markets in mexico: Definition, databases, and descriptive analysis.

Alix-Garcia, J. and Sellars, E. A. (2020). Locational Fundamentals, Trade, and the Changing Urban Landscape of Mexico. *Journal of Urban Economics*, 116:103213.

Andersson, F., Burgess, S., and Lane, J. I. (2007). Cities, Matching and the Productivity Gains of Agglomeration. *Journal of Urban Economics*, 61:112–128.

Bacolod, M., Blum, B. S., Rangel, M. A., and Strange, W. C. (2023). Learners in cities: Agglomeration and the spatial division of cognition. *Regional Science and Urban Economics*, 98:103838.

Banco de México (2020). Informe Trimestral, Julio-Septiembre 2021.

Banco de México (2023). Informe Trimestral, Julio-Septiembre 2023.

Bassier, I. (2023). Firms and Inequality When Unemployment is High. *Journal of Development Economics*, 161:103029.

Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., and Zhang, Q. (2017). Roads, railroads, and decentralization of chinese cities. *The Review of Economics and Statistics*, 99(3):435–448.

Baum-Snow, N. and Pavan, R. (2012). Understanding the City Size Wage Gap. *The Review of Economic Studies*, 79(1):88–127.

Behrens, K., Duranton, G., and Robert-Nicoud, F. (2014). Productive Cities: Sorting, Selection, and Agglomeration. *Journal of Political Economy*, 122(3):507–553.

Bleakley, H. and Lin, J. (2012). Thick-market effects and churning in the labor market: Evidence from us cities. *Journal of urban economics*, 72(2-3):87–103.

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, 87(3):699–739.

Cadena, B. (2014). Recent immigrants as labor market arbitrageurs: Evidence from the minimum wage. *Journal of Urban Economics*, 80(C):1–12.

Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018). Firms and Labor Market Inequality: Evidence and Some Theory. *Journal of Labor Economics*, 36(S1):S13–S70.

Card, D., Heining, J., and Kline, P. (2013). Workplace Heterogeneity and the Rise of West German Wage Inequality. *The Quarterly Journal of Economics*, 128(3):967–1015.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024a). Binscatter Regressions.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Feng, Y. (2024b). On Binscatter. *American Economic Review*, 114(5):1488–1514.

Chauvin, J. P., Glaeser, E., Ma, Y., and Tobio, K. (2017). What is Different About Urbanization in Rich and Poor Countries? Cities in Brazil, China, India and the United States. *Journal of Urban Economics*, 98:17–49.

Combes, P.-P., Démurger, S., Li, S., and Wang, J. (2020). Unequal Migration and Urbanisation Gains in China. *Journal of Development Economics*, 142:102328.

Combes, P.-P. and Gobillon, L. (2015). The Empirics of Agglomeration Economies. In *Handbook of regional and urban economics*, volume 5, pages 247–348. Elsevier.

Dauth, W., Findeisen, S., Moretti, E., and Suedekum, J. (2022). Matching in Cities. *Journal of the European Economic Association*, 20(4):1478–1521.

D'Costa, S. and Overman, H. G. (2014). The Urban Wage Growth Premium: Sorting or Learning? *Regional Science and Urban Economics*, 48:168–179.

De la Roca, J., Parkhomenko, A., and Velásquez-Cabrera, D. (2023). Skill Allocation and Urban Amenities in the Developing World. Working paper.

De la Roca, J. and Puga, D. (2016). Learning by Working in Big Cities. *The Review of Economic Studies*, 84(1):106–142.

Diallo, Y., Sarr, I., and Diagne, I. (2022). Role of Firms in Wage Dispersion: Evidence from a Developing Country. *PEDL Research Papers*.

Duranton, G. (2015). Growing through cities in developing countries. *World Bank Research Observer*, 30(1):39–73.

Duranton, G. (2016). Agglomeration Effects in Colombia. *Journal of Regional Science*, 56(2):210–238.

Duranton, G. and Puga, D. (2004). Micro-foundations of Urban Agglomeration Economies. In *Handbook of regional and urban economics*, volume 4, pages 2063–2117. Elsevier.

Eckert, F., Hejlesen, M., and Walsh, C. (2022). The return to big-city experience: Evidence from refugees in denmark. *Journal of Urban Economics*, 130:103454.

Fowler, C. S. and Jensen, L. (2020). Bridging the Gap Between Geographic Concept and the Data We Have: The Case of Labor Markets in the USA. *Environment and Planning A: Economy and Space*, 52(7):1395–1414.

Frías, J. A., Kaplan, D. S., Verhoogen, E., and Alfaro-Serrano, D. (2022). Exports and Wage Premia: Evidence from Mexican Employer-Employee Data. *The Review of Economics and Statistics*, pages 1–45.

Ghani, E., Goswami, A. G., and Kerr, W. R. (2016). Highway to success: The impact of the golden quadrilateral project for the location and performance of indian manufacturing. *The Economic Journal*, 126(591):317–357.

Gould, E. D. (2007). Cities, Workers, and Wages: A Structural Analysis of the Urban Wage Premium. *The Review of Economic Studies*, 74(2):477–506.

Grover, A., Lall, S., and Timmis, J. (2023). Agglomeration economies in developing countries: A meta-analysis. *Regional Science and Urban Economics*, 101:103901.

Helsley, R. W. and Strange, W. C. (1990). Matching and agglomeration economies in a system of cities. *Regional Science and urban economics*, 20(2):189–212.

Henderson, J. V. (1986). Efficiency of resource usage and city size. *Journal of Urban economics*, 19(1):47–70.

Jedwab, R., Ianchovichina, E., and Haslop, F. (2022). Consumption cities versus production cities: new considerations and evidence. Policy Research Working Paper WPS 10105, World Bank Group, Washington, D.C.

Jochmans, K. and Weidner, M. (2019). Fixed-effect regressions on network data. *Econometrica*, 87(5):1543–1560.

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

Kumler, T., Verhoogen, E., and Frías, J. (2020). Enlisting employees in improving payroll tax compliance: Evidence from mexico. *Review of Economics and Statistics*, 102(5):881–896.

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

Millimet, D. L. and Bellemare, M. F. (2025). On the (Mis)Use of the Fixed Effects Estimator. Working paper.

Moretti, E. and Yi, M. (2024). Size matters: Matching externalities and the advantages of large labor markets. Working Paper 32250, National Bureau of Economic Research.

Pérez Pérez, J. and Nuño-Ledesma, J. G. (2024). Workers, workplaces, sorting, and wage dispersion in mexico. *Economía LACEA Journal*, 23(1).

Petrongolo, B. and Pissarides, C. (2006). Scale effects in markets with search. *The Economic Journal*, 116(508):21–44.

Puggioni, D., Calderón, M., Cebreros Zurita, A., Fernández Bujanda, L., Inguanzo González, J. A., and Jaume, D. (2022). Inequality, Income Dynamics, and Worker Transitions: The case of Mexico. *Quantitative Economics*, 13(4):1669–1705.

Rotemberg, J. J. and Saloner, G. (2000). Competition and human capital accumulation: a theory of interregional specialization and trade. *Regional Science and Urban Economics*, 30(4):373–404.

Ulyssea, G. (2010). Regulation of Entry, Labor Market Institutions and the Informal Sector. *Journal of Development Economics*, 91(1):87–99.

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

Wheeler, C. H. (2008). Local market scale and the pattern of job changes among young men. *Regional Science and Urban Economics*, 38(2):101–118.

World Bank (2023). Employment in services (% of total employment). Accessed: 2023-06-10.

Online Appendix - Not for Publication

A Data and Descriptive Statistics

We use IMSS data previously analyzed in Pérez Pérez and Nuño-Ledesma (2024). IMSS refers to the Instituto Mexicano del Seguro Social, a Mexican government agency responsible for public health, pension management, and social security. Private sector employees are required by law to register with IMSS. According to Mexico's labor survey, about 83% of the formal workforce in 2022 were registered with IMSS. Self-employed individuals have the option to register with IMSS, granting them access to certain aspects of the social security system. Typically, self-employed workers register with the equivalent of one legal minimum salary. Self-employed records comprise approximately 0.1% of the overall IMSS database. In cases where a worker reports multiple jobs, we retain the job with the highest reported wage. Only 2.5% of workers reported having more than one job in December 2018.

The IMSS social security information is available every month. We analyze records spanning from November 2004 to December 2018. Our analysis concludes in 2018 due to significant changes in Mexico's labor market resulting from the COVID-19 pandemic and substantial increases in the minimum wage from 2019 to 2022. The database initially consists of 12.8 million workers in November 2004 and grows to 20.1 million workers by December 2018. Our primary variable of interest is the daily taxable income, which encompasses various forms of compensation, excluding additional non-taxable payments such as paid vacations and bonuses. Wages exceeding 25 UMAs (units of measure and update) are capped. About 1,539 pesos per day, or approximately 80 USD at 2018 exchange rates. We rely on the *registro patronal* as our workplace identifier, as opposed to alternatives like the *Registro Federal de Contribuyentes* (RFC), because *registros* better capture employer-worker relationships in settings where multiple employers operate within the same plant. The RFC is an identification number issued by the Mexican Tax Administration Service to individuals, firms, and any other legal entities for tax compliance purposes. The *registro patronal* is an employer registry identification number assigned to all employers with the purpose of managing pensions and the provision of other social security benefits through the Mexican Social Security Institute. The *registro patronal* identifiers are anonymized by our data provider before access is granted. The procedure used to mask the identifiers is inconsequential to our analysis and

does not affect our econometric analysis. Because the *registro patronal* is employer-based rather than plant-based, our results reflect employer effects rather than the plant effects traditionally reported in literature using AKM models (e.g., Card et al. 2013).

Table A.1, reproduced from Pérez Pérez and Nuño-Ledesma (2024), displays a summary of the IMSS data. We divide our data into three time intervals: 2004-2008, 2009-2013, and 2014-2018. Within each year, our sample includes a substantial number of wage observations, ranging from 73 to 113 million for men aged 25-54. In column (2) of the table, we observe a 0.7% decrease in the average real daily wage for prime-age men between 2009 and 2014 when compared to 2005. However, this decline is followed by a 1.5% increase by 2018. Column (3) illustrates a widening gap in earned wages between 2005 and 2018.

Table A.1: Descriptive Statistics: Prime-age Men, National Level

		Real wage		
	(1)	(2)	(3)	(4)
	Observations	Mean	Std. dev	Percent censored
2005	73,847,545	394.589	406.167	2.675
2009	80,065,916	394.602	402.992	2.690
2014	96,354,574	394.200	409.212	2.649
2018	110,844,774	401.186	412.367	2.058

Source: Authors' calculations using IMSS data. Observations correspond to the sum of all the monthly observations in a year. Real wages are daily taxable income registered in IMSS, expressed in real terms using prices from July 2018. The percent censored is the percentage of observations with wages exactly equal to the upper wage limit of 25 minimum wages or units of measure, which are updated per year.

Abowd et al. (1999) show that AKM estimates identify fixed effects for workers and workplaces within a "connected set" of workplaces where there is a shared pool of workers who switch jobs at least once. Our estimates use the largest connected set within each time interval. A workplace is part of the connected set if at least one of its workers has worked or will work in a different workplace during the given time interval. Direct connections between every pair of workplaces are not necessary for a connected set to exist.

Table A.2, reproduced from Pérez Pérez and Nuño-Ledesma (2024), shows the number of worker-month observations for prime-age men that had more than one job, the number of individuals, and the average and standard deviation of log wages. In each interval, our database comprises

324 to 518 million worker-month observations, representing 12 to 16 million individuals. The standard deviation of salaries slightly increased from 0.81 in the 2004-2008 interval to 0.83 in the 2014-2018 interval. Average real wages exhibited a consistent upward trend throughout the sample. Columns (5) to (8) of Table A.2 display the corresponding descriptive statistics for the largest connected set of prime-age male workers. The largest connected set encompasses at least 96% of all worker-year observations and 96% of all individuals within a given interval. Average wages within the connected set are slightly higher than those in the overall sample, while standard deviations are marginally smaller. Given the substantial size of the connected set relative to the entire sample, the similar mean salaries, standard deviations, and comparable trends in average wages and salary dispersion, our focus on this connected group does not involve a significant loss of detail.

Table A.2: Descriptive Statistics - Overall Sample and Workers in the Largest Connected Set

	All sample				Individuals in largest connected set			
			Log wage				Log wage	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Interval	All obs.	Individuals	Mean	Std. dev.	All obs.	Individuals	Mean	Std. dev.
Nov 2004-2008	324,468,447	11,835,313	5.627	0.813	311,941,032	11,363,073	5.657	0.808
Ratio: largest connected/all					96.14	96.01	100.53	99.39
2009-2013	431,227,399	13,526,466	5.600	0.826	417,008,147	13,083,589	5.625	0.823
Ratio: largest connected/all					96.70	96.73	100.45	99.65
2014-2018	518,128,252	15,920,775	5.609	0.831	505,015,793	15,512,438	5.628	0.829
Ratio: largest connected/all					97.47	97.44	100.35	99.71
Change from first to last interval			-0.018	0.018			-0.029	0.021

Source: Authors' calculations using IMSS data. Statistics for men aged 25 to 54 years who held more than one job, i.e., were employed in more than one workplace during the analysis period. Log wage is the log of daily taxable income registered in IMSS, expressed in real terms using prices from July 2018. "Ratio: largest connected/all" is the ratio of the corresponding statistic in the largest connected set to its counterpart in the full sample.

We complement the IMSS data with commuting-zone level and city-level covariates. At the commuting-zone level, we calculate the informality rate and the mean years of schooling with municipality data from the 2005, 2015, and 2020 Population and Housing Censuses. At the metro level, we calculate unionization rates using the Occupation and Employment Survey of the National Institute of Statistics and Geography (INEGI). The informality rate is the percentage of workers who do not have social security benefits over the total number of workers. The unionization rate is the percentage of formal workers who state that they belong to a union, divided by the total number of formal workers. We calculate quarterly rates from 2004 to 2018 for the 43 largest Mexican cities and average rates at the city level for estimation periods (2004-2008, 2009-2013, 2014-2018).

B Imputation of censored wages

According to the IMSS database documentation, daily earnings were censored at 25 daily minimum wages before February 2017 and at 25 "units of measurement and actualization" (UMA) thereafter. On average, these values amount to 1,539 pesos per day, which corresponds to approximately 80 USD at 2018 exchange rates. The censoring may impact our estimates of the variance of wages and the fraction of this variance attributed to positive assortative matching. Therefore, for our main analysis, we impute wage values for these top-coded observations.

While wages should be top-coded at either 25 minimum wages or 25 UMAs, in practice, we find some wage values greater than these censoring limits. Therefore, to impute daily earnings for top-coded observations, we examined the right tail of the wage distribution starting at the documented censoring threshold. Specifically, we plotted the wage density and searched for the most frequent value immediately above the theoretical censoring point. We interpret this modal spike as the actual censoring point. This empirical approach reflects the idea that, under censoring, a large number of observations will accumulate at the top-coding threshold. To account for potential reporting errors or rounding issues, we allowed for a 5-peso margin above the documented value when identifying the mode.

After identifying the censoring thresholds, we followed the methodology in Card et al. (2013) to impute censored wages. Specifically, we estimate Tobit models for wages where we regress them on a set of independent variables and obtain predictions for the wages in the top-coded ob-

servations. The independent variables in each model include the worker's age; the number of employees in the firm; a dummy indicating whether the firm has 10 or more employees; the average wage in the firm; the fraction of employees in the firm with censored wages; the fraction of times the worker's wage is censored in the sample; and a dummy indicating if the worker is observed in only one month.

Estimating the Tobit model in our individual-level data proved unfeasible. We note, however, that the estimates of the Tobit model coefficients, except for the variance of the residuals, can be obtained from an estimation on a cell-aggregated dataset using frequency weights. Therefore, we estimate two Tobit regressions: a first one using a 1% random sample of workers, including the independent variables previously mentioned; and a second one aggregating the data by period, region, industry, workers observed only once in the database, and age group. [25]

We estimate over 169 Tobit regressions using aggregated data, one for each month from November 2004 to December 2018 [26] The separate estimation allows model coefficients and censoring thresholds to vary over time, ensuring that predicted values reflect time-specific wage structures. With the estimated coefficients from the second Tobit model and the variance of the residuals retrieved from the worker-level Tobit model in the 1% sample, we apply the formula proposed by Card et al. (2013) to estimate the uncensored log wage y^u for each censored observation:

$$y^u = X'\hat{\beta} + \hat{\sigma}\Phi^{-1}\left[k + u \times (1-k)\right], \tag{B.1}$$

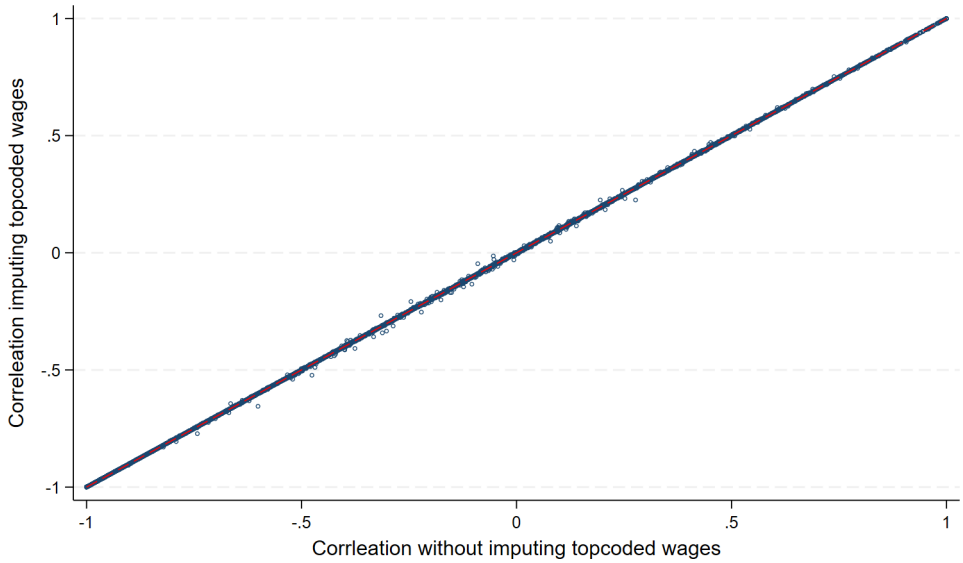
Here, the variable y^u is the imputed uncensored log wage, $X'\beta$ is the predicted mean from the Tobit model, $\hat{\sigma}$ is the estimated standard deviation of the residuals of the Tobit model, the function Φ^{-1} is the inverse CDF of the standard normal distribution, and the variable $k = \Phi\left(\frac{c-X'\beta}{\sigma}\right)$ represents the CDF evaluated at the standardized censoring point c. Additionally, the variable u is a random draw from a uniform distribution. This approach enables us to generate plausible values for censored wages by leveraging the distributional assumptions of the Tobit model and introducing randomness through u, thereby ensuring variability in the imputed values.

[25] This approach follows the methodology proposed by Card et al. (2013), who estimate separate Tobit models on cells defined by individual and firm characteristics. We are unable to fully replicate their cell construction since we lack information on some firm-level educational characteristics, such as the mean years of schooling and the fraction of university graduates at the current firm.

[26] We follow Card et al. (2013), who estimate a series of Tobit models by year.

After imputing wages for the censored workers, we obtain the correlation between worker and firm fixed effects from AKM estimates using data that include the imputed wages. We then compared the correlation between firm and worker fixed effects before and after the imputation. Figure B.1 shows that correlations at the commuting zone-industry level remained stable, suggesting that the imputation does not meaningfully alter our estimates.

Figure B.1: Comparison of Correlations with and without Imputation of Censored Daily Earnings



Source: Authors' calculations using IMSS data.

C Details about Commuting Zone Construction

We use commuting zones for Mexico calculated by Aldeco et al. (2024). We provide a summary of their methodology for building commuting zones here. Using residence-workplace data from the Mexican census of 2010, Aldeco et al. (2024) group municipalities according to their similarity in commuting patterns, using the same methodology as Fowler and Jensen (2020) for the US. First, they calculate a commuting dissimilarity index for each pair of municipalities i, j, using data on commuting flows f_{ij} from an origin-destination survey:

$$D_{ij} = 1 - \frac{f_{ij} + f_{ji}}{\min\left(\sum_l f_{il}, \sum_l f_{lj}\right)} \tag{C.2}$$

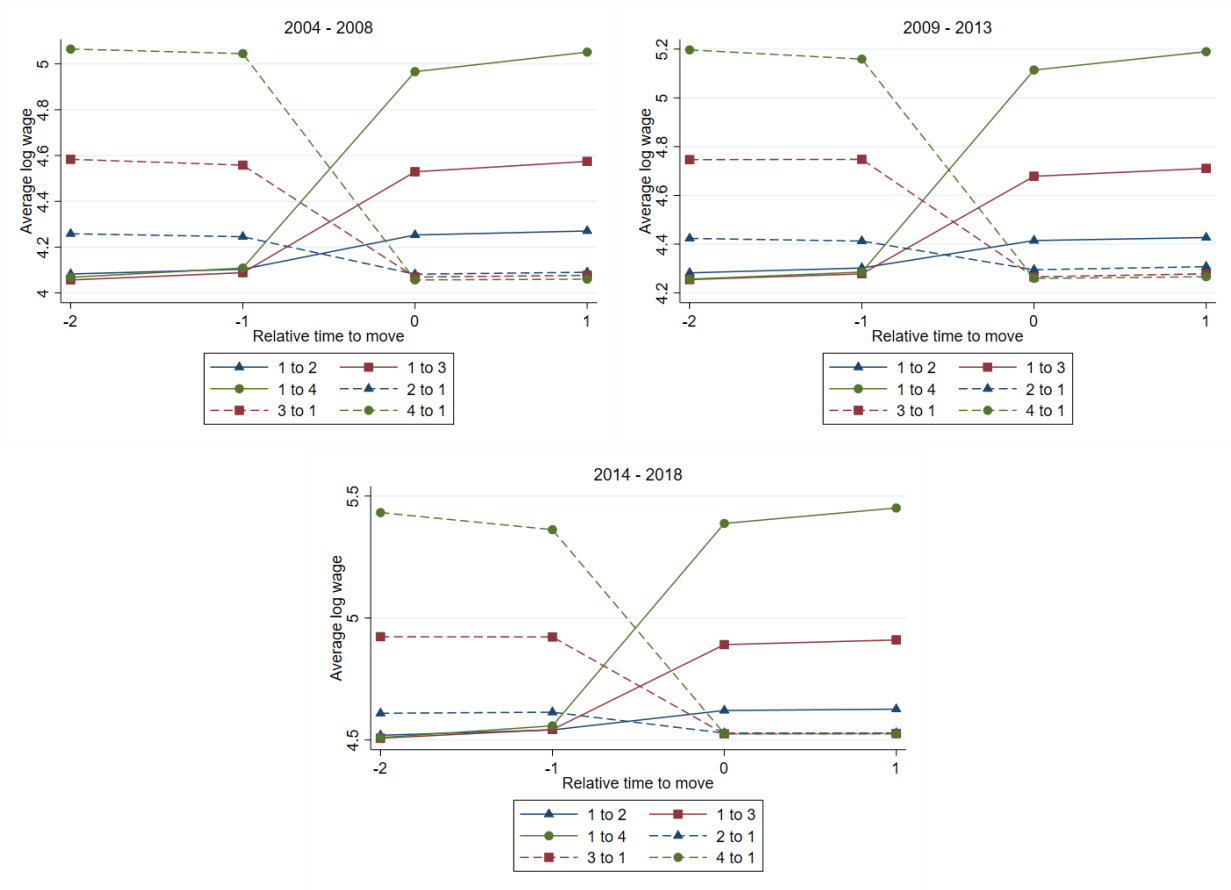
This index grows larger as the share of workers who commute between municipalities i and j becomes smaller. After building the index, they cluster the municipalities based on this index using a hierarchical clustering algorithm.

D Exchangeability

We reproduce the evidence of exchangeability shown in Pérez Pérez and Nuño-Ledesma (2024) in this section. According to Card et al. (2013), if the residual term in equation (1) is uncorrelated with the variables on the right-hand side, workers who move from workplace A to workplace B should, on average, experience a wage change opposite in sign to workers moving in the opposite direction. To explore this in our dataset, we follow Card et al. (2013) and present an event study in Figure D.1, adapted from Pérez Pérez and Nuño-Ledesma (2024). The plot illustrates the average wages of workers who changed jobs during each time interval of our analysis period. These workers may transition from "low-wage" to "high-wage" workplaces or vice versa. We categorize workplaces based on the quartile of the average wage of their co-workers in the initial job and the corresponding quartile in the final job. We then calculate average wages before and after the job switch for each category. Our analysis excludes observations from establishments with only one worker and focuses on "direct" moves, which are defined as moves without an unemployment spell between jobs.

The Figure D.1 reveals that different mobility groups, classified by the average wage of co-workers, have distinct average wage levels before and after a job move. Before the move, average wages in the quartile of origin exhibit a monotonic variation with respect to the destination quartile. For instance, workers moving from quartile 4 (the highest average co-worker wage) to quartile 1 (the lowest mean co-worker wage) have higher average wages before the job switch compared to those moving from quartile 3 to quartile 1, and so on. Additionally, the absolute change in average salaries when transitioning from one quartile to another is equivalent in magnitude to the variation associated with the opposite change. This symmetry aligns with an additive wage model that incorporates worker and workplace fixed effects, similar to the one we estimate.

Figure D.1: Exchangeability: Average Log Wage Around Movement by Quartile of Average co-workers' Wages in the Origin and Destination Workplace



Source: Authors' calculations using IMSS data. The graph shows the average wages of workers who move between an origin workplace and a destination workplace, from two months before the move to one month after the move. The lines group workers according to the quartiles of average co-worker wages in the origin and destination workplaces. The panels correspond to different time intervals. We exclude observations from establishments with only one worker. We retain only "direct" moves without an unemployment spell in the transition between jobs.

E Variance Decomposition with Limited Mobility Bias Corrections

In this section, we demonstrate that the results of variance decompositions are similar in the baseline estimates and those with limited mobility bias corrections. Table E.1 presents variance decompositions with various corrections.

Columns 1 to 3 show the baseline estimates. Columns 4 to 6 display estimates from a model that clusters firms into groups, following Bonhomme et al. (2019). We calculate twenty percentiles of the within-firm distribution of wages. Then, we cluster firms into five clusters using the percentile values as variables for clustering. Last, we reestimate the AKM model using firm cluster indicators instead of firm indicators and recalculate the variance decomposition. The results show slightly higher variance shares attributed to assortative matching.

Columns 7 to 9 show results of the variance decomposition using a leave-one-out variance components estimator from Kline et al. (2020). In this case, the variance shares are similar to those from the baseline estimates.

Table E.1: Variance Decompositions with Different Corrections for Limited Mobility Bias

	Baseline			Bonhomme et al. (2019)			Kline et al. (2020)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Interval 1 2004-2008	Interval 2 2009-2013	Interval 3 2014-2018	Interval 1 2004-2008	Interval 2 2009-2013	Interval 3 2014-2018	Interval 1 2004-2008	Interval 2 2009-2013	Interval 3 2014-2018
Variance and covariance									
Total variance of log wages	0.65	0.68	0.69	0.65	0.68	0.69	0.64	0.67	0.67
Variance of worker effects	0.29	0.27	0.25	0.28	0.26	0.25	0.29	0.27	0.25
Variance of workplace effects	0.21	0.24	0.25	0.16	0.19	0.20	0.20	0.23	0.25
2 Cov(worker effects, workplace effects)	0.10	0.12	0.13	0.16	0.17	0.18	0.10	0.12	0.13
Variance shares									
Variance of worker effects	0.44	0.40	0.37	0.43	0.39	0.37	0.46	0.41	0.37
Variance of workplace effects	0.33	0.36	0.37	0.25	0.28	0.29	0.32	0.35	0.37
2 Cov(person effects, workplace effects)	0.16	0.18	0.20	0.25	0.25	0.26	0.17	0.19	0.21

Source: Authors' calculations using IMSS data. The columns display the results of variance decompositions following equation (2) of the main text. Columns (1) to (3) display the baseline estimates using unadjusted estimated worker and workplace fixed effects. Columns (4) to (6) display estimates using the Bonhomme et al. (2019) correction, where workplaces are clustered into five clusters according to within-workplace wage distributions before estimating the AKM model. Columns (7) to (9) show estimates using the Kline et al. (2020) correction, where the variance components are calculated using leave-one-out estimators over the connected set with observations from January, May and September for each year.

F Additional Estimates

Table F.1: Estimates with Log Employment as Independent Variable

	(1) CZ	(2) CZ-Industry
Dependent variable: correlation of worker and workplace FE		
Log employment	0.0500***	0.0621***
	(0.003)	(0.003)
R^2	0.245	0.153
N	1,961	10,118
Mean of dep. variable	-0.014	0.000

Source: Author's calculations using IMSS data. The columns display the results of regressions of the correlation coefficient between worker and workplace effects from AKM model estimates and log employment, at the commuting zone and commuting zone industry levels. All the regressions pool data from the three intervals: 2004-2008, 2009-2013, and 2014-2018, and include dummies by interval. CZ stands for commuting zone. For column (2), we restrict to cells with more than 50 workers and more than 5 firms. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

Table F.2: City Size and Assortative Matching in Mexico's Formal Labor Markets: Estimates by Interval

Dependent variable: correlation of worker and workplace FE		
	(1)	(2)
	CZ	CZ-Industry
A: 2004-2008		
Log Population	0.0689***	0.0569***
	(0.007)	(0.004)
N	614	3,209
R^2	0.147	0.085
Mean of dep. variable	-0.036	-0.025
B: 2009-2013		
Log Population	0.0663***	0.0554***
	(0.007)	(0.004)
N	668	3,387
R^2	0.130	0.081
Mean of dep. variable	-0.033	-0.013
C: 2014-2018		
Log Population	0.0685***	0.0477***
	(0.006)	(0.003)
N	679	3,522
R^2	0.166	0.073
Mean of dep. variable	0.024	0.037

Source: Author's calculations using IMSS and INEGI data. The columns display the results of regressions of the correlation coefficient between worker and workplace effects from AKM model estimates and log employment, at the commuting zone and commuting zone industry levels. Each panel corresponds to a different time interval. CZ stands for commuting zone. For column (2), we restrict to cells with more than 50 workers and more than five firms. Clustered standard errors at the commuting-zone level in parentheses. *: $p<0.1$, **: $p<0.05$, ***: $p<0.01$.

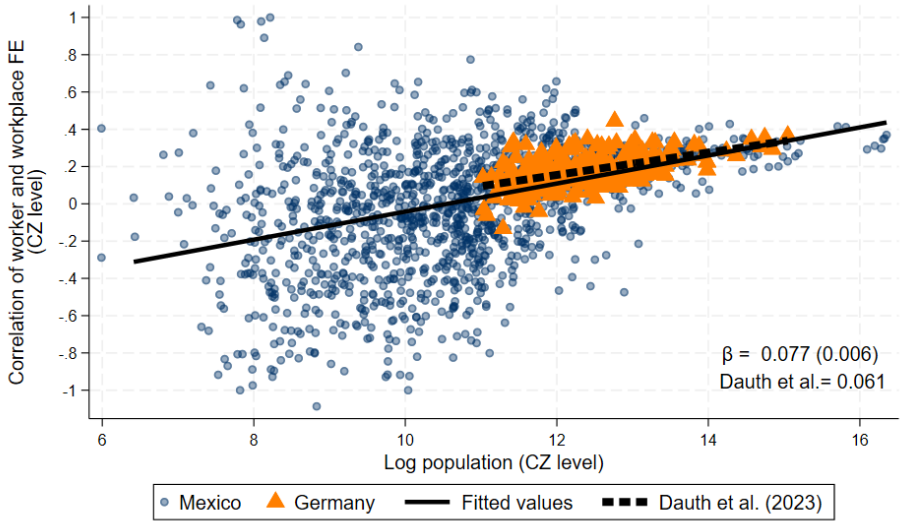
Table F.3: City Size and Assortive Matching in Mexico's Formal Labor Markets: Heteroskedasticity adjustment.

Dependent variable: correlation of worker and workplace FE		
	(1)	(2)
	CZ	CZ-Industry
A: Baseline Model		
Log Population	0.0679***	0.0531***
	(0.004)	(0.004)
B: Weighted Least Squares (non-parametric weights)		
Log Population	0.0605***	0.0506***
	(0.003)	(0.003)
N	1,961	10,118
Mean of dep. variable	-0.014	0.000

Source: Author's calculations using IMSS and INEGI data. This table addresses heteroskedasticity concerns through WLS following the agnostic approach in Cadena (2014). Specifically, we estimate the baseline model and then regress the squared residuals of this baseline regression on cell size with a local linear regression model. We then re-estimate the baseline regression, weighting each observation by the inverse of the variance conditional on cell size as predicted by the local linear regression model. We iterate this process three times. We use an Epanechnikov kernel for the local-linear regression with the bandwidth determined by minimizing the MSE of the prediction. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

G Counterfactual estimates with Germany's industrial composition and city size distribution

Figure G.1: City Size and Assortative Matching in Mexico's Formal Labor Markets. Germany's industrial composition counterfactual at the commuting zone level.



Source: Author's calculations using IMSS, INEGI, and Eurostat data. The figure displays a scatterplot illustrating the relationship between log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone level. We reweighted the sample to match Germany's industrial composition and re-estimated the correlation between worker and workplace fixed effects. For comparison, the figure displays the relationship estimated for Germany by Dauth et al. (2022). The bottom-right values indicate the slope of a linear regression corresponding to the displayed relationship. The regression includes dummies for each time interval. Clustered standard errors at the commuting zone level in parentheses.

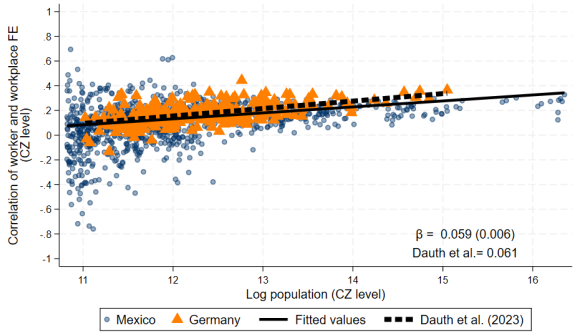
Table G.1: City Size and Assortative Matching in Mexico's Formal Labor Markets. Germany's industrial composition counterfactual.

Dependent variable: correlation of worker and workplace FE	
	(1)
	CZ
A: Germany's industrial composition counterfactual	
Log Population	0.0773***
	(0.006)
R^2	0.155
N	1,333
Mean of dep. variable	-0.013

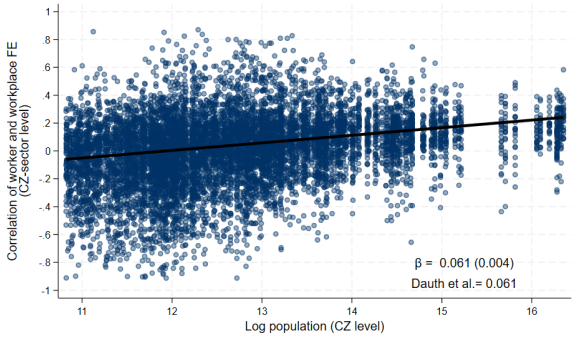
Source: Author's calculations using IMSS, INEGI, and Eurostat data. "CZ" stands for commuting zone, respectively. Regressions pool data from 2009-2013 and 2014-2018 with interval dummies. We reweighted the sample to match Germany's industrial composition and re-estimated the correlation between worker and workplace fixed effects. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

Figure G.2: City Size and Assortative Matching in Mexico's Formal Labor Markets. Excluding commuting zones with less than 50,000 inhabitants.

(a) Commuting Zone



(b) Commuting Zone - Industry



Source: Author's calculations using IMSS and INEGI data. Each panel displays a scatterplot illustrating the relationship between log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone and commuting zone industry levels. We exclude CZs with less than 50,000 inhabitants. For comparison, panel (a) displays the relationship estimated for Germany by Dauth et al. (2022). For panel (b), we restrict to cells with more than five firms and more than 50 workers. We classify industries according to a 2-digit NAICS classification. The bottom-right values display the slope of a linear regression corresponding to the displayed relationship. The regressions include dummies for each time interval. Clustered standard errors at the commuting-zone level in parentheses.

Table G.2: City Size and Assortative Matching in Mexico's Formal Labor Markets. Excluding commuting zones with less than 50,000 inhabitants.

Dependent variable: correlation of worker and workplace FE		
	(1)	(2)
	CZ	CZ-Industry
A: Baseline Model		
Log Population	0.0589***	0.0614***
	(0.006)	(0.004)
R^2	0.167	0.094
B: Correlation of worker and residual workplace FE		
Log Population	0.0572***	0.0612***
	(0.006)	(0.004)
R^2	0.181	0.093
C: Log population instrumented with population in 1921-1950		
Log Population	0.0400	0.0523***
	(0.029)	(0.019)
R^2	0.154	0.092
First-stage F	4.181	3.001
N for Panels A-C	877	9,174
Mean of dep. variable	0.092	0.009
D: Corrected for limited mobility bias		
Log Population	0.0454***	0.0456***
	(0.006)	(0.003)
R^2	0.117	0.069
N	877	9,163
Mean of dep. variable	0.258	0.190
E: Dropping the 10% largest areas		
Log Population	0.0594***	0.0605***
	(0.007)	(0.005)
R^2	0.163	0.075
N	874	7,897
Mean of dep. variable	0.091	0.024

Source: Author's calculations using IMSS and INEGI data. "CZ" stands for commuting zone. Regressions pool data from 2004-2008, 2009-2013, and 2014-2018 with interval dummies. The estimations exclude CZs with less than 50,000 inhabitants. Column (2) restricts to cells with over five firms and 50 workers. Panels: A - baseline; B - industry-demeaned workplace fixed effects; C - we instrument population with historical population at the Metro and CZ levels, relying on historical population estimates from Alix-Garcia and Sellars. Log population instrumented with log population in 1921, 1930, 1940; and 1950 ; D - Bonhomme et al. (2019)'s limited-mobility bias correction with five workplace clusters, and E - excluding extreme populations. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

H Details about Historical Population Estimates and Additional Instrumental-Variable Results

For Panel C in Table 2, we instrument the population with the historical population at the commuting zone (CZ) level. We use Alix-Garcia and Sellars' (2020) historical population estimates to calculate Mexico's historical population levels. Alix-Garcia and Sellars (2020) divide Mexico's territory into 15x15 km grid cells and estimate the population in each cell from 1900 to 1950. We intersect these grid cells with contemporary municipality boundaries and calculate the historical population of a municipality m as the sum of the populations of each cell that intersects with municipality m, weighted by the proportion of municipality m's land that intersects each cell. Finally, we aggregate these estimates to the commuting-zone level.

Table H.1: City Size and Assortative Matching in Mexico's Formal Labor Markets Instrumenting Current Population with Historical Population

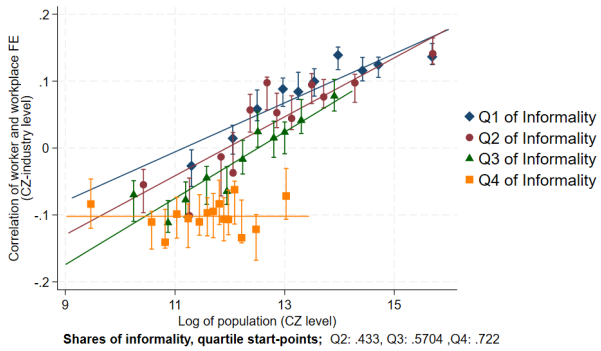
Dependent variable: correlation of worker and workplace FE		
	(1)	(2)
	CZ	CZ-Industry
A: Baseline Model		
Log Population	0.0679***	0.0531***
	(0.004)	(0.004)
R^2	0.155	0.090
B: Log population instrumented with population in 1921-1950		
Log Population	0.0592***	0.0420***
	(0.007)	(0.013)
R^2	0.152	0.086
First-stage F	174.786	4.757
Hansen J statistic	5.651	2.663
p-value Hansen J stat	0.129	0.446
C: Log population instrumented with population in 1930-1950		
Log Population	0.0594***	0.0411***
	(0.009)	(0.013)
R^2	0.153	0.086
First-stage F	232.884	6.290
Hansen J statistic	4.468	2.663
p-value Hansen J stat	0.107	0.264
D: Log population instrumented with population in 1940-1950		
Log Population	0.0588***	0.0432***
	(0.009)	(0.014)
R^2	0.152	0.087
First-stage F	350.603	4.541
Hansen J statistic	0.553	2.175
p-value Hansen J stat	0.457	0.140
N	1,961	10,118
Mean of dep. variable	-0.014	0.000

Source: Author's calculations using data from IMSS, INEGI, and Alix-Garcia and Sellars (2020). The table presents the results of instrumental variable regressions of the correlation coefficient between worker and workplace effects from AKM model estimates and log population at the commuting zone and commuting zone-industry levels. We use historical population estimates from Alix-Garcia and Sellars to calculate Mexico's historical population at the commuting-zone level. See appendix section H for details on the construction of historical population estimates. All the regressions pool data from the three intervals: 2004-2008, 2009-2013, and 2014-2018, and include dummies by interval. CZ stands for commuting zone. For column 4, we restrict to cells with more than five firms and more than 50 workers. Panel A shows baseline estimates. First-stage F is the first-stage F-statistic. Clustered standard errors at the commuting-zone level in parentheses. *: p<0.1, **: p<0.05, ***: p<0.01.

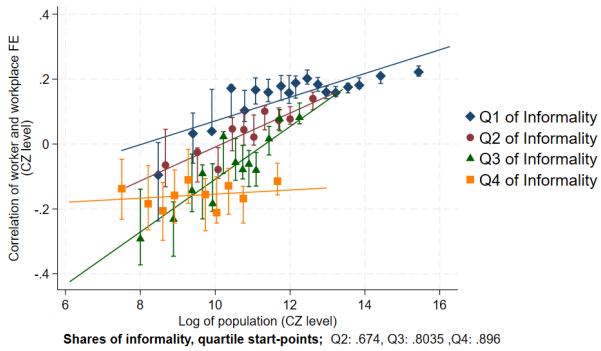
I Additional Results on Mechanisms Affecting Matching Externalities

Figure I.1: City Size and Assortative Matching in Mexico's Formal Labor Markets by Informality Quartiles.

(a) Commuting Zone-Industry



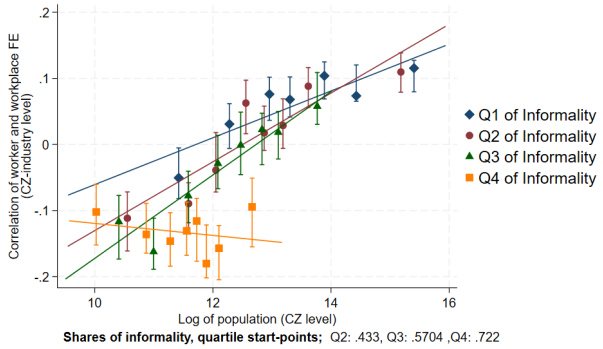
(b) Commuting Zone



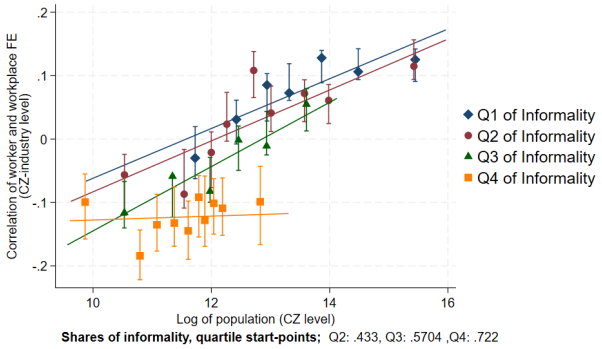
Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry and commuting zone levels for each quartile of the informality rate. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

Figure I.2: City Size and Assortative Matching in Mexico's Formal Labor Markets. Commuting Zone-Industry by Informality Quartiles. Estimates by Interval

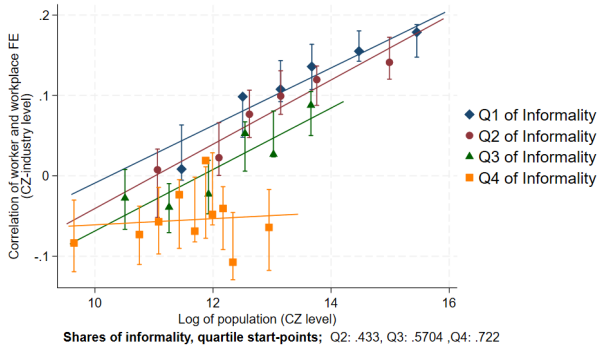
(a) 2004-2008



(b) 2009-2013



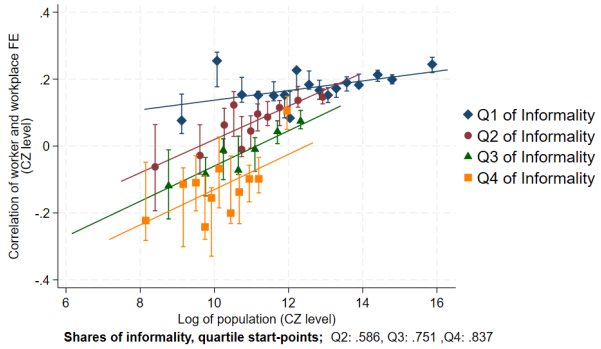
(c) 2014-2018



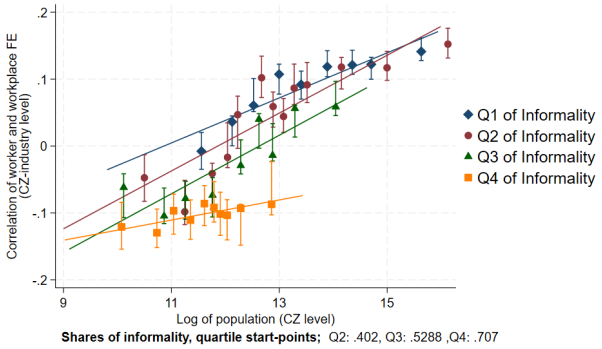
Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry level for each quartile of the informality rate. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plot. The panels correspond to different time intervals.

Figure I.3: City Size and Assortative Matching in Mexico's Formal Labor Markets by Informality Quartiles Excluding the Southern Region.

(a) Commuting Zone-Industry



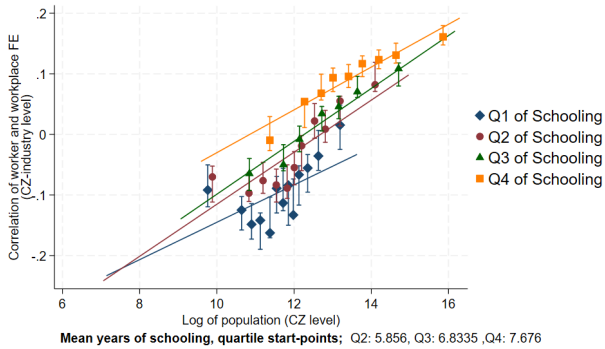
(b) Commuting Zone



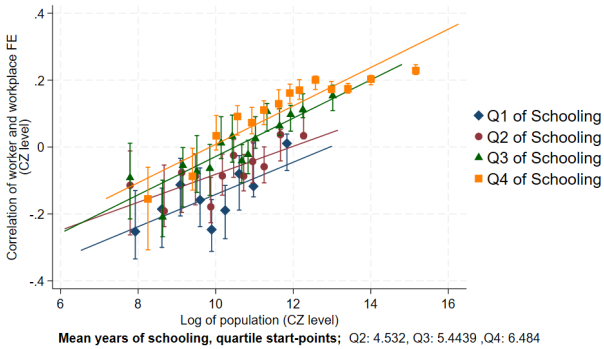
Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry and commuting zone levels for each quartile of the informality rate. The sample excludes commuting zones in Mexico's southern region, which comprises the following states: Campeche, Chiapas, Guerrero, Oaxaca, Quintana Roo, Tabasco, Veracruz, and Yucatán. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

Figure I.4: City Size and Assortative Matching in Mexico's Formal Labor Markets by Years of Schooling Quartiles.

(a) Commuting Zone-Industry



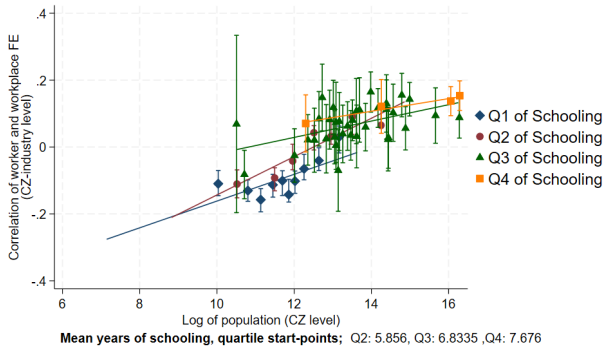
(b) Commuting Zone



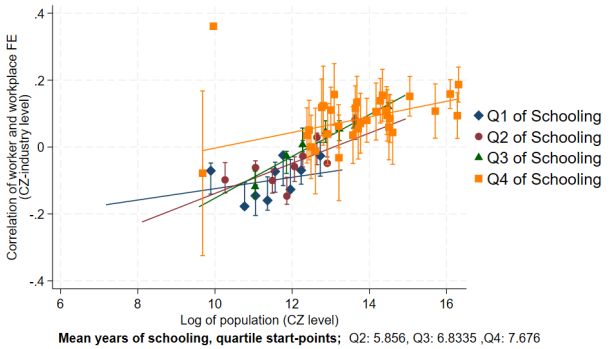
Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry and commuting zone levels for each quartile of the mean years of schooling. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

Figure I.5: City Size and Assortative Matching in Mexico's Formal Labor Markets. Commuting Zone-Industry by Years of Schooling Quartiles. Estimates by Interval

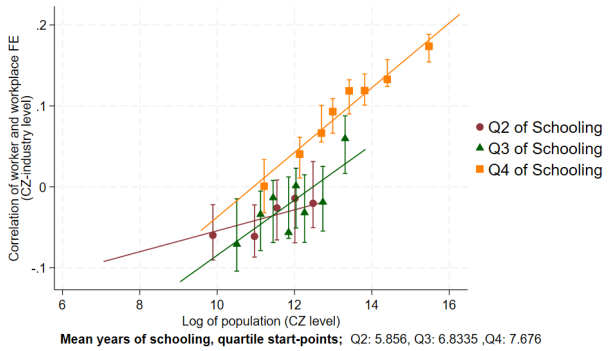
(a) 2004-2008



(b) 2009-2013



(c) 2014-2018



Source: Author's calculations using IMSS and INEGI data. The figure displays a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the commuting zone-industry level for each quartile of the mean years of schooling. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots. The panels correspond to different time intervals. Panel (c) excludes the first schooling quartile due to an insufficient effective sample size for estimating the binscatter.

J Union Coverage

The study of unionization as a potential limiting factor for positive assortative matching in labor markets dates back to Card et al. (2013), who attribute an increase in the relevance of assortative matching in wage inequality in Germany to a decline in union power. Unions may prevent the rotation and dismissal of union-affiliated workers, limiting assortative matching.

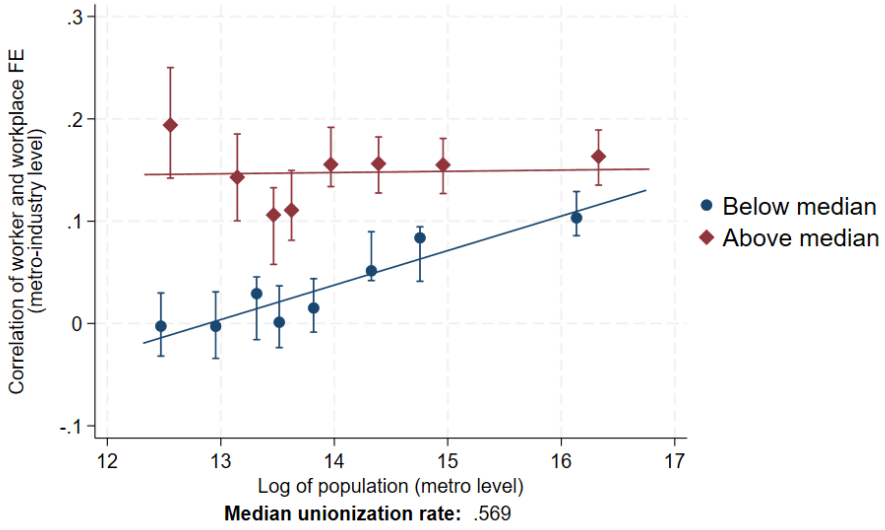
Due to data limitations, unionization rates are only available at the metro-industry level, rather than at our preferred geographical unit of analysis (CZ and CZ-industry level). As a result, these results are not directly comparable to those obtained for mechanisms such as schooling and informality. Nevertheless, they offer complementary evidence on the role of institutional factors—such as union presence—in shaping assortative matching dynamics across local labor markets.

We explore the relationship between unionization rates and assortative matching in local labor markets in Figure J.1. We split the sample at the median unionization rate (56.9%). Our findings are mixed. In areas with below-median unionization, the strength of positive assortative matching is lower than in areas with above-median unionization. On the other hand, in metropolitan areas with below-median unionization, we observe a positive relationship between matching and city size: assortative matching improves with local labor market size. In contrast, for metropolitan areas with above-median unionization rates, this relationship becomes less pronounced.

This pattern suggests two possible interpretations. First, strong unions may institutionalize wage and employment norms that substitute for market-based matching, resulting in hiring practices that are less dependent on productivity alignment. Second, relatively rigid labor institutions may constrain the flexibility needed to exploit agglomeration effects, muting the advantages of urban scale.

These results are consistent with the idea that strong unions flatten the relationship between city size and sorting, either by enforcing non-market job allocation rules or by limiting firm and worker discretion during hiring decisions.

Figure J.1: City Size and Assortative Matching in Mexico's Formal Labor Markets. Metropolitan Area-Industry and Unionization Rate



Source: Author's calculations using IMSS and INEGI data. The figure shows a binned scatter plot of the log population and the correlation between estimated worker and workplace effects from AKM models at the metro-industry level for units above and below the median unionization rate.. The vertical bars are confidence intervals for the conditional mean of the correlation at each level of (log) population. We used the `binsreg` and `binstest` commands (Cattaneo et al., 2024a,b) with default settings to generate the scatter plots.

Appendix References

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, 67(2):251–333

Aldeco, L., Calderón, M., Chiquiar, D., Hanson, G., Pérez Pérez, J., and Velázquez, C. (2024). Local labor markets in mexico: Definition, databases, and descriptive analysis

Alix-Garcia, J. and Sellars, E. A. (2020). Locational Fundamentals, Trade, and the Changing Urban Landscape of Mexico. *Journal of Urban Economics*, 116:103213

Bonhomme, S., Lamadon, T., and Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, 87(3):699–739

Card, D., Heining, J., and Kline, P. (2013). Workplace Heterogeneity and the Rise of West German Wage Inequality. *The Quarterly Journal of Economics*, 128(3):967–1015

Cadena, B. (2014). Recent immigrants as labor market arbitrageurs: Evidence from the minimum wage. *Journal of Urban Economics*, 80(C):1–12

Fowler, C. S. and Jensen, L. (2020). Bridging the Gap Between Geographic Concept and the Data We Have: The Case of Labor Markets in the USA. *Environment and Planning A: Economy and Space*, 52(7):1395–1414

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

Pérez Pérez, J. and Nuño-Ledesma, J. G. (2024). Workers, workplaces, sorting, and wage dispersion in mexico. *Economía LACEA Journal*, 23(1)